

Publication

# PISA 2022 Technical Report





# PISA 2022 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 Member countries of the OECD.

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.

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.

Note by the Republic of Türkiye

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. Türkiye recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Türkiye 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 Türkiye. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

**Please cite this publication as:**

OECD (2024), *PISA 2022 Technical Report*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/01820d6d-en>.

ISBN 978-92-64-92890-9 (print)

ISBN 978-92-64-82476-8 (PDF)

PISA

ISSN 1990-8539 (print)

ISSN 1996-3777 (online)

**Photo credits:** Cover © Monkey Business Images/Shutterstock.com.

Corrigenda to OECD publications may be found at: <https://www.oecd.org/en/publications/support/corrigenda.html>.

© OECD 2024

---

This work is available under the [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO](https://creativecommons.org/licenses/by-nc-sa/3.0/) (CC BY-NC-SA 3.0 IGO). For specific information regarding the scope and terms of the licence as well as possible commercial use of this work or the use of PISA data please consult Terms and Conditions on [www.oecd.org](https://www.oecd.org).

---

# Foreword

The OECD Programme for International Student Assessment (PISA) assesses the extent to which 15-year-old students near the end of compulsory education have acquired the knowledge and skills that are essential for full participation in modern societies. Eighty-one countries and economies took part in its eighth cycle, PISA 2022, in which a comprehensive set of indicators, spanning student performance, attitudes towards learning, school environment and resources, and many other aspects of school life were collected for comparison and analysis in what is the world's largest comparative education study.

PISA findings have a significant impact on education systems worldwide, so it is essential that its data is collected within rigorous technical standards and following the best practices in educational assessment to assure valid, reliable, and internationally comparable findings.

This Technical Report aims to describe and provide clarity on the technical and methodological underpinnings of these findings. Its goal is twofold: to act as a tool for quality control to enable all stakeholders involved in PISA, as well as the general public to evaluate the quality of the data released in PISA 2022 regarding its four guiding principles of validity, reliability, comparability and fairness outlined in its Technical Standards; and to empower data analysts and researchers to understand this study and its outputs, thus fostering the further use of PISA data.

The PISA results were achieved through the work of PISA International Contractors that led the study at the international level, National Centres led by National Project Managers that implemented PISA in each participating country and economy, and subject matter expert groups. Chapter 1 of this Technical Report describes in the different groups involved in the implementation of PISA 2022, and Annex J contains a list of individuals who contributed to this PISA cycle.

The publication was co-ordinated by the OECD Secretariat. Tiago Fragoso co-ordinated the production of the report with Eugenio Gonzalez from Educational Testing Service (ETS) and with support from Juliana Andrea González Rodríguez. Contributions were received from Francesco Avvisati, Natalie Foster, Tue Halgreen, Miyako Ikeda, and from the ETS, ACT, Westat and cApStAn teams listed in Annex J. Charlotte Baer, Eda Cabbar, and Della Shin provided communications assistance, and Thomas Marwood, Valeria Pelosi, and Ricardo Sanchez Torres provided editorial and administrative support.

This revised edition of the PISA 2022 technical report expands upon the initial March 2024 release, featuring updates to chapters 6, 14, and 18, integrating new insights on Creative Thinking and Financial Literacy.

# Table of contents

Foreword	3
<b>1 Programme for International Student Assessment - An Overview</b>	<b>12</b>
Introduction	12
Participation	14
Features of PISA	14
Technical innovations in PISA 2022	15
Managing and implementing PISA	17
PISA 2022 publications	18
References	18
Annex 1.A. Participation in PISA 2022	23
<b>2 The PISA 2022 Integrated Assessment Design</b>	<b>24</b>
Introduction	24
PISA 2022 integrated design	24
Overview of the field trial assessment design	26
References	43
Notes	43
Annex 2.A. Main survey items	44
Annex 2.B. Core testlet	50
Annex 2.C. Adaptive testlet	52
<b>3 Test Development for the Core Domains</b>	<b>55</b>
Introduction	55
The 2022 mathematics assessment framework	57
Role of the mathematics expert group in item development	63
PISA 2022 test development	63
Field trial	67
Main survey	70
References	75
Annex 3.A. Test developments for the core domain	76
<b>4 Creative Thinking Test Design and Test Development</b>	<b>79</b>
Introduction	79
The Role of the Creative Thinking Expert Group in Item Development	79
PISA 2022 Creative Thinking Assessment Framework	80
PISA 2022 Innovative Domain Assessment Design	81
PISA 2022 Innovative Domain Assessment Development	82

Field Trial	83
Main Survey	86
Preparation of data collection instruments	86
References	87
Notes	88
Annex 4.A. Creative thinking items	89
<b>5 Context Questionnaire Development</b>	<b>91</b>
Introduction	91
The role of the PISA context questionnaire framework in development	91
Questionnaires for different respondent groups	93
Phases of Questionnaire Development and QUALITY ASSURANCE	94
Summary	97
References	98
Notes	98
Annex 5.A. PISA 2022 Questionnaire Framework Content Modules	99
Annex 5.B. Student Questionnaire	100
<b>6 Sample Design</b>	<b>103</b>
Target population and overview of the sampling design	103
Population coverage, and school and student participation rate standards	104
Main survey school sample	109
School samples	119
Student samples	123
Teacher samples	125
Definition of school	126
References	126
Notes	126
Annex 6.A. Sample design	127
<b>7 Translation and Verification of the Survey Material</b>	<b>133</b>
Introduction	133
PISA Countries/economies, Languages, Scope and Verifier training	134
Materials subject to verification	134
Verifier qualifications, training and instructional materials	136
Testing languages and translation/adaptation procedures	138
Development of source versions	139
Double translation from two source languages	143
Training and instructional materials for national translation teams	144
Centralised trend material transfer	145
Questionnaire adaptation negotiation	146
International verification of the national versions	147
Main survey verification	159
Annex 7.A. Translation items	165
Notes	171
Annex 7.B. Verifier interventions	172
Annex 7.C. Translatability assessment items	173
Annex 7.D. Additional items	174

<b>8 Field Operations</b>	<b>175</b>
Overview of roles and responsibilities	175
The selection of the school sample	178
Preparation of school-level materials	178
The selection of the student sample	178
Packaging and shipping materials	179
Test administration	180
Receipt of materials at the national centre after testing	182
Field Trial and Main Survey reviews	182
Notes	183
<b>9 PISA Quality Monitoring</b>	<b>184</b>
Introduction	184
Field Trial and Main Survey review questionnaires	184
National centre consultations	185
PISA Quality Monitor Hiring Process	185
PISA Quality Monitor Training	185
PISA Quality Monitor Visits	186
Data adjudication	187
<b>10 Survey Weighting and the Calculation of Sampling Variance</b>	<b>188</b>
Survey weighting	188
Teacher weighting	194
Calculating sampling variance	196
References	199
Annex 10.A. School non-response items	201
<b>11 Scaling PISA data</b>	<b>206</b>
Overview	206
Data yield and data quality	207
IRT modelling and scaling	218
Population modelling and multiple imputation	223
Analysis of data with plausible values	225
Application to the PISA 2022 Main Survey	227
References	238
Notes	242
Annex 11.A. Detailed Procedures and Techniques	243
<b>12 Data Management Procedures</b>	<b>251</b>
Introduction	251
Data management at the international and national level	252
The data management process and quality control	253
Preparing files for public use and analysis	263
Notes	265
Annex 12.A. Additional Data Management Items	266
<b>13 Sampling Outcomes</b>	<b>275</b>
Population coverage	275
Teacher response rates	277
Design effects and effective sample sizes	278



Variability of the design effect	280
References	282
Annex 13.A. Sampling outcomes	283
<b>14 Scaling outcomes</b>	<b>291</b>
IRT scaling outcomes	291
Population modelling outcomes	298
Linking error	300
References	302
Notes	302
Annex 14.A. IRT Scaling Outcomes and Population Modelling Analysis	303
<b>15 Coding Design, Coding Process, and Reliability Studies</b>	<b>313</b>
Introduction	313
Coding design	313
Coding preparation	316
Coding procedures	318
Reliability studies	320
Machine-supported coding system	323
References	324
Annex 15.A. Detailed Overview of the Coding Process	325
<b>16 Data Adjudication</b>	<b>331</b>
Introduction	331
General outcomes	335
Notes	341
Annex 16.A. Data adjudication additional items	342
<b>17 Proficiency Scale Construction for the Core Domains</b>	<b>346</b>
Introduction	346
Development of the PISA scales	348
Defining the proficiency levels	350
Reporting PISA results for Mathematics	353
Defining levels of mathematical literacy	353
Cutpoints defining proficiency levels for Reading, Science and Financial Literacy in PISA 2022	354
References	354
Annex 17.A. PISA Mathematics reporting scales	355
<b>18 PISA 2022 Innovative Domain Test Design and Test Development</b>	<b>364</b>
Introduction	364
The role of the Creative Thinking Expert Group (CTEG) in the framework and item development	364
PISA 2022 creative thinking assessment framework	365
PISA 2022 innovative domain test assembly design	370
PISA 2022 innovative domain assessment design and development	370
Field trial	373
PISA 2022 main survey	377
Data adjudication and approach to scaling the data for reporting	378
References	381
Annex 18.A. Development and Validation of the Creative Thinking Assessment in PISA 2022	382

<b>19 Scaling procedures and construct validation of context questionnaire data</b>	<b>384</b>
Introduction	384
Scaling methodology and reporting of scores	386
Student Questionnaire derived variables	392
Financial Literacy Questionnaire derived variables	409
ICT Familiarity Questionnaire derived variables	412
Well-Being Questionnaire derived variables	415
Parent Questionnaire derived variables	417
School Questionnaire derived variables	419
Teacher Questionnaire derived variables	432
References	440
Annex 19.A. Methodology and Overview of Derived Variables in PISA 2022 Context Questionnaires	443
<b>20 Questionnaire Design and the Computer-Based Questionnaire Platform</b>	<b>448</b>
Introduction	448
Questionnaire Design	448
Student-Administered Questionnaires	449
School Questionnaire	452
Teacher Questionnaire	453
Parent Questionnaire	454
Computer-based Questionnaire Platform	455
General questionnaire development process	472
Overview of the technical infrastructure	481
Summary	481
Annex 20.A. Evolution and Implementation of Questionnaire Administration in PISA 2022	482
<b>21 The PISA 2022 Computer-based Platform</b>	<b>484</b>
Introduction	484
Item rendering	484
Translation and online item review	485
School computer requirements	486
System diagnostic	486
Test delivery system	486
Data capture and scoring student responses	488
Open-ended coding system	489
<b>22 International data products</b>	<b>490</b>
Public-use files	490
Codebooks for the PISA 2022 public-use data files	492
Data analysis and software tools	492
Population and quality check of the PISA Data Explorer	494
IEA's International Database Analyzer	495
Notes	497

Annex A. Item Pool Classification tables	498
Annex B. Contrast Coding Tables	499
Annex C. Student and School Sample Size Tables	500
Annex D. National Household Possession Items Tables	501
Annex E. Final Distribution of RMSD Values Across Groups for Each Scale Tables	504
Annex F. Common and Unique Item Parameters in Each Domain, by Country and Language Tables	505
Annex G. Equated P Tables	506
Annex H. Testing Periods Tables	507
Annex I. PISA 2022 Technical Standards and guidelines	508
Purpose of document	508
Format of the document	510
Scope	510
Data standards	510
Management standards	522
National involvement standards	527
Definitions	528
Annex J. PISA 2022 Contractors, Staff and Consultants	532
PISA Governing Board	533
PISA 2022 National Project Managers	536
OECD Secretariat	538
Mathematics Expert Group (MEG)	540
Extended Mathematics Expert Groups (eMEG)	540
Financial Literacy Expert Group (FLEG)	540
Creative Thinking Expert Group (CTEG)	541
Questionnaire Expert Group (QEG)	541
Questionnaire Senior Framework Advisors	541
ICT Expert Group	542
Technical Advisory Group	542
PISA 2022 Lead Contractors	542
PISA 2022 Contributors, working with Lead Contractors	546

## FIGURES

Figure 2.1. Overview of the PISA 2022 main survey integrated design	31
Figure 2.2. Overview of main survey computer-based MSAT design - with creative thinking and without creative thinking	35
Figure 2.3. Overview of the hybrid main survey computer-based MSAT design for mathematics	36
Figure 2.4. Example testlet structure across stages for one group	37
Figure 2.5. Average relative efficiency of MSAT paths over linear forms for PISA 2022 mathematics test design	42

Figure 3.1. Mathematical literacy for PISA 2022	58
Figure 3.2. Cognitive processes and the mathematical problem-solving model: prior to 2022 (left) and for 2022 (right)	59
Figure 4.1. Competency model for the PISA test of creative thinking	81
Figure 5.1. PISA 2022 Questionnaire Framework and Modules	92
Figure 5.2. Virtual and in-person meetings with the PISA 2022 Questionnaire Expert Group (QEG)	93
Figure 6.1. School response rate standards	107
Figure 7.1. Translation workflow for the production of a French source version of newly-developed PISA 2022 Mathematics units	141
Figure 7.2. Trend Transfer process diagram	146
Figure 7.3. Sample of a test adaptation spreadsheet (TAS) from the PISA 2022 Field Trial	148
Figure 7.4. Verification workflow of trend items	149
Figure 7.5. Distribution by category of verifier interventions in New Mathematics units (translated versions)	155
Figure 7.6. Number of issues per national version in New Mathematics units (translated versions)	156
Figure 7.7. Distribution by category of verifier interventions in Creative Thinking units (translated versions)	156
Figure 7.8. Number of issues per national version in Creative Thinking units (translated versions)	157
Figure 7.9. Distribution by category of verifier interventions in New Mathematics units (adapted versions)	158
Figure 7.10. Number of issues per national version in New Mathematics units (adapted versions)	158
Figure 7.11. Distribution by category of verifier interventions in Creative Thinking units (adapted versions)	159
Figure 7.12. Number of issues per national version in Creative Thinking units (adapted versions)	159
Figure 11.1. Main sample yield for countries/economies participating in the CBA	208
Figure 11.2. Financial literacy sample yield for participating countries/economies	209
Figure 11.3. Main sample yield for countries/economies participating in the PBA and new PBA	209
Figure 11.4. Mathematics median response time by median proficiency across countries/economies	214
Figure 11.5. Distribution of mathematics response time in each country/economy	215
Figure 11.6. Item response curve (ICC) for an item where the common item parameter is not appropriate for one group	221
Figure 11.7. Residual correlation matrix for the creative thinking main survey	231
Figure 11.8. Percentage of variance from principal component analyses for 6 countries/economies	232
Figure 12.1. Overview of the data management process	253
Figure 14.1. Frequency of invariant, variant, and dropped items for mathematics, by country/economy	293
Figure 14.2. Frequency of invariant, variant, and dropped items for reading, by country/economy	293
Figure 14.3. Frequency of invariant, variant, and dropped items for science, by country/economy	294
Figure 14.4. Frequency of invariant, variant, and dropped items for creative thinking, by country/economy	294
Figure 14.5. Frequency of invariant, variant, and dropped items for financial literacy, by country/economy	295
Figure 14.6. Distribution of the first plausible values and item RP62 values in mathematics	296
Figure 14.7. Distribution of the first plausible values and item RP62 values in reading	297
Figure 14.8. Distribution of the first plausible values and item RP62 values in science	297
Figure 14.9. Distribution of the first plausible values and item RP62 values in creative thinking	298
Figure 14.10. Distribution of the first plausible values and item RP62 values in financial literacy	298
Figure 15.1. Organization of multiple coding for the CBA designs	315
Figure 15.2. Organization of multiple coding for the PBA and New PBA standard coding design	316
Figure 17.1. Simplified relationship between items and students on a proficiency scale	347
Figure 17.2. Calculating the RP values used to define PISA proficiency levels	352
Figure 18.1. The PISA 2022 competency model for creative thinking	366
Figure 18.2. General coding process for 'generate diverse ideas' items	368
Figure 18.3. General coding process for 'generate creative ideas' and 'evaluate and improve' items	369
Figure 19.1. Types of derived variables for questionnaires in PISA 2022	385
Figure 19.2. Category characteristic curves for a four-category item under the generalised partial credit model (GPCM)	388
Figure 19.3. Illustration of how an increase in the slope parameter $\alpha$ affects the category characteristic curves of the model above	389
Figure 19.4. Computation of ESCS in PISA 2022	407
Figure 20.1. Questionnaire Authoring Tool home page	456
Figure 20.2. Questionnaire Authoring Tool: Main View (with a specific question example)	457
Figure 20.3. Questionnaire Authoring Tool: Organisation of Main View	457
Figure 20.4. The expanded view information	458
Figure 20.5. Preview of a question in the QAT	459
Figure 20.6. Upload XLIFF for Preview feature in the QAT	460
Figure 20.7. Upload XLIFF for Preview – Edit Columns feature	461

Figure 20.8. Information Template	462
Figure 20.9. Exclusive Choice Template	463
Figure 20.10. Multiple-Choice Template	464
Figure 20.11. List of Exclusive Choice (Table) Template	464
Figure 20.12. List of Multiple Choice (Table) Template	465
Figure 20.13. List of Text Inputs Template	466
Figure 20.14. Multiple List of Text Inputs (Table) Template	466
Figure 20.15. Scale Question Type Template	467
Figure 20.16. Free Text Input Template	468
Figure 20.17. Drop-Down Template	469
Figure 20.18. Drop-Down (Table) Template	470
Figure 20.19. Consistency Check Rule Template	470
Figure 20.20. Consistency Check Message	471
Figure 20.21. Routing Rule Template	471
Figure 20.22. Question IDs	472
Figure 20.23. PISA 2022 computer-based questionnaire life cycle	474
Figure 20.24. Translation of Questionnaires into multiple national languages	476
Figure 20.25. Questionnaire Platform – Administrative View	477
Figure 20.26. Distribution of the PISA 2022 servers	479
Figure 22.1. PISA database population and quality control	494

## TABLES

Table 2.1. Field trial computer-based assessment design	29
Table 2.2. Field trial paper-based assessment designs	30
Table 2.3. Main survey paper-based assessment designs	32
Table 2.4. Main survey computer-based assessment design	34
Table 2.5. Main survey computer-based UH form design	38
Table 2.6. Field trial computer-based financial literacy design	39
Table 2.7. Main survey computer-based financial literacy design	40
Table 7.1. Sample of a questionnaire adaptation spreadsheet (QAS) from the PISA 2022 Field Trial	151
Table 7.2. Main Survey Questionnaire Change Request Form in the QAS	161
Table 7.3. Overview of Testing and Questionnaire Items	164
Table 8.1. Timing of the CBA and PBA assessment sessions	181
Table 11.1. Detailed analysis: Scaling PISA data	237
Table 14.1. Global Analysis of Item Dynamics and Student Proficiency in PISA 2022	301
Table 20.1. Field trial computer-based design for Student Questionnaire	450
Table 20.2. Field Trial Paper-based Design for Students	451
Table 20.3. Field trial computer-based design for Teacher Questionnaires (TCQ)	454

## BOXES

Box 1.1. Key features of PISA 2022	15
Box 6.1. Illustration of probability proportional to size (PPS) sampling	113

# 1 Programme for International Student Assessment - An Overview

## Introduction

The OECD Programme for International Student Assessment (PISA) is a collaborative effort among OECD Member countries and non-Member partner 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/economies in 2001 and 2002, respectively), the second in 2003, the third in 2006, the fourth in 2009 (followed by a further 10 countries/economies in 2010), the fifth in 2012, the sixth in 2015, the seventh in 2018, and the eighth in 2022. The results of these surveys have been published in a series of reports (OECD, 2020<sup>[1]</sup>; 2020<sup>[2]</sup>; 2020<sup>[3]</sup>; 2019<sup>[4]</sup>; 2019<sup>[5]</sup>; 2019<sup>[6]</sup>); (OECD, 2017<sup>[7]</sup>; 2017<sup>[8]</sup>; 2017<sup>[9]</sup>; 2016<sup>[10]</sup>; 2016<sup>[11]</sup>); (OECD, 2014<sup>[12]</sup>; 2014<sup>[13]</sup>; 2014<sup>[14]</sup>; 2013<sup>[15]</sup>; 2013<sup>[16]</sup>); (OECD, 2011<sup>[17]</sup>; 2010<sup>[18]</sup>; 2010<sup>[19]</sup>; 2010<sup>[20]</sup>); (OECD, 2010<sup>[21]</sup>; 2010<sup>[22]</sup>; 2007<sup>[23]</sup>; 2004<sup>[24]</sup>; 2001<sup>[25]</sup>); (OECD/UNESCO Institute for Statistics, 2003<sup>[26]</sup>; Walker, 2011<sup>[27]</sup>) and a wide range of thematic and technical reports, e.g. OECD (OECD, 2021<sup>[28]</sup>; 2021<sup>[29]</sup>).

The next survey will occur in 2025. For each assessment, reading, mathematics or science is chosen as the major domain and given greater emphasis than the remaining two domains. In 2000, 2009, and 2018, the major domain was reading; in 2003, 2012, and 2022 it was mathematics, in 2006 and 2015 it was science as it will be in 2025.

The three-year cadence of PISA cycles was disrupted by the coronavirus (COVID-19) global pandemic along with education systems worldwide. The implementation of the PISA 2021 Field Trial was impacted by the first wave of school closures and the uncertainty of when and how schools would reopen led the PISA Governing Board (PGB) to decide on postponing the ongoing PISA 2021 cycle and the upcoming PISA 2024 cycles by one year. Both cycles were renamed PISA 2022 and PISA 2025, respectively, and are thus referred throughout this report for coherence.

PISA is an age-based survey, assessing 15-year-olds in school in grade 7 or higher. These students are approaching the end of compulsory schooling in most participating countries/economies, and school enrolment at this level is close to universal in most 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 mathematical knowledge and skills to solve various kinds of numerical and spatial challenges

and problems; 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; 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 2022 domains are fully described in the *PISA 2022 Assessment and Analytical Framework* (OECD, 2023<sup>[30]</sup>).

PISA also conducts assessments of additional cross-curricular competencies from time to time as participating countries/economies 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 using a computer-delivered platform, which was followed for its third administration in PISA 2018. In PISA 2022 financial literacy was assessed for the fourth time, also using a computer-based platform, and was administered to 20 countries/economies. A computer-based assessment of critical thinking was also added in PISA 2022 and administered to 65 countries/economies.

In addition, PISA administers Student Questionnaires to collect information from students on various aspects of their home, family and school background, and School Questionnaires to collect information from school principals about various aspects of organisation and educational provision in schools. Both background questionnaires also included the PISA 2022 Global Crises Module (Bertling et al., 2020<sup>[31]</sup>), developed to measure several aspects of the disruption caused by the school closures during the COVID-19 pandemic to students, and measures taken by schools. There are also optional questionnaire modules for students asking about Familiarity with Information and Communications Technology (ICT) and Well-being (WB).

In PISA 2022, 17 countries/economies also administered a Parent Questionnaire to the parents of the students participating in PISA. A Teacher Questionnaire was implemented in PISA 2018 and was administered in 19 countries/economies. In PISA 2022, a Student Well-being Questionnaire was also administered in 15 countries.

Annex Table 1.A.2. 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/economies 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/economies;
- differences in the proportion of variation in achievement between (rather than within) schools, and differences in this value across countries/economies;
- differences between countries/economies 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/economies;
- changes in any or all of these relationships over time by linking the current and previous PISA cycles.

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 that describes the PISA 2022 questionnaires is included in the *PISA 2022 Assessment and Analytical Framework* (OECD, 2023<sup>[30]</sup>).

## Participation

The first PISA survey was implemented in 43 countries/economies (including 32 OECD Member countries). It was first conducted in 2000 in 32 countries/economies (including 28 OECD Member countries) using written tasks answered in schools under independently supervised test conditions. Another 11 countries/economies completed the same assessment in 2001 and 2002. PISA 2000 surveyed reading, mathematics, and science with a primary focus on reading.

The following cycle took place in 2003 with a focus in mathematics, in 2006 with a focus on science and every three years since then, including an increasing number of OECD Member countries, Associates, and Partner countries and economies. A detailed account of participation in PISA since 2000 can be found in Annex Table 1.A.2. The eighth cycle of PISA, was originally scheduled to take place in 2021, but it was postponed by one year, from 2021 to 2022, due to the COVID-19 pandemic. This cycle was renamed PISA 2022 and it covered reading, mathematics, science, creative thinking, and financial literacy. Mathematics was its primary focus and was implemented in 37 OECD countries and 45 partner countries/economies.

The participants in PISA 2022 are listed in Annex Table 1.A.2. The figure also indicates whether countries/economies participated in the computer-based (CBA) or paper-based mode (PBA), and shows the countries/economies that participated in the critical thinking (CrT) and/or financial literacy assessment.

## Features of PISA

The technical characteristics of the PISA survey involve several 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/economies, and taking into account the diverse cultural contexts of those countries/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/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 2022. Box 1.1 provides an overview of the central design elements of PISA 2022.



### Box 1.1. Key features of PISA 2022

#### The content

The PISA 2022 survey focused on mathematics, with reading, science as minor areas of the assessment, and creative thinking as an innovative domain. PISA 2022 also included an assessment of young people's financial literacy, which was optional for participating 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

Some 690 000 students completed the assessment in 2022, representing about 29 million 15-year-olds in the schools of the 81 participating countries/economies.

#### The assessment

Computer-based tests were used in most countries, with assessments lasting a total of two hours. In mathematics and reading, a multi-stage adaptive approach was applied in computer-based tests whereby students were assigned a block of test items based on their performance in preceding blocks.

Test items were a mixture of multiple-choice questions and questions requiring students to construct their own responses. The items were organised into groups based on a passage of text describing a real-life situation. More than 15 hours of test items for reading, mathematics, science and creative thinking were covered, with different students taking different combinations of test items.

Students also answered a background questionnaire, which took about 35 minutes to complete. The questionnaire sought information about the students themselves, their attitudes, dispositions and beliefs, their homes, and their school and learning experiences. School principals completed a questionnaire that covered school management and organisation, and the learning environment. Both students and schools responded to the Global Crises Module additional items in their respective questionnaires, assessing how school closures caused by the COVID-19 pandemic affected student lives and school policies.

Some countries/economies also distributed additional questionnaires to elicit more information. These included: in 19 countries/economies, a questionnaire for teachers asking about themselves and their teaching practices; and in 17 countries/economies, a questionnaire for parents asking them to provide information about their perceptions of and involvement in their child's school and learning.

Countries/economies could also choose to distribute three other optional questionnaires for students: 53 countries/economies distributed a questionnaire about students' familiarity with computers and 15 countries/economies distributed a questionnaire about students' well-being.

## Technical innovations in PISA 2022

PISA 2015 represented the first step of switching from a primarily paper-based survey that included optional computer-based modules to a fully computer-delivered survey, a process that continued into the 2018 and was further expanded into the 2022 cycle. The computer-based delivery mode allows PISA to measure new and expanded aspects of the domain constructs. In mathematics, new material was

incorporated aimed to move away from the need to perform basic calculations to assess mathematical reasoning and its interplay with problem solving. PISA 2022 extended and improved the computer-based multi-stage adaptive testing design implemented for the reading literacy domain in PISA 2018 to the assessment of mathematical literacy in the 2022 cycle, further improving measurement accuracy and efficiency, especially at the extremes of the proficiency distribution. In financial literacy, in PISA 2018 some interactive tasks were created that allowed students to manipulate variables and observe effects of financial choices. These were also included in PISA 2022. In addition, in PISA 2022, new tasks were created to fill in gaps in the framework coverage left by previously released tasks. Additionally, PISA 2022 retained a paper-based version of the assessment that included only trend units. This paper-based assessment was administered in a small number of countries/economies that did not implement the computer-based survey (see Annex Table 1.A.2.). Chapter 2 describes the integrated assessment design, and Chapter 20 describes the technical aspects of the computer delivery platform. Chapter 19 describes the platform used for the development and delivery of background questionnaires for students, school principals and teachers.

In addition to the implementation of PISA 2022 as a fully computer-based survey, an interactive portal was further developed to support survey implementation and enhance communication between national teams and the international contractors. Throughout this report references are made to the PISA Portal as it was used in a variety of tasks during the implementation of PISA 2022.

Roll-out of on- online marking of tests continued in PISA 2022 following its successful adoption as the main medium of test marking in PISA 2018. This mode 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 manual data entry, saving time and resources, and avoiding potential operator error. Chapter 15 describes the marking process while Chapter 20 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 20 describes the type of information that was collected.

The innovations in the scaling model implemented in 2015 continued in 2022 to improve the measurement of trends across PISA cycles. The ability to establish and maintain trends over time is an important goal for PISA. The integrated design for the assessment which is described in Chapter 2 further expanded on the 2018 design by increasing 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 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 2022, as do 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 from PISA 2015 onwards 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 response data, PISA 2015 and later cycles implemented a hybrid model which combined a Rasch approach with a two-parameter-logistic model and a generalised partial credit model (GPCM) used where appropriate. Chapter 11 describes this innovative approach in detail and Chapter 14 presents scaling outcomes.

## Managing and implementing PISA

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

Experts from participating countries/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 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/economies ensured that the instruments were internationally valid, that they took the cultural and educational contexts of participating countries/economies into account, that the assessment materials had strong measurement potential, and that the instruments emphasised authenticity and educational validity. See Annex J for the list of members of the expert groups.

Each of the participating country/economy appointed a National Project Manager (NPM) to implement PISA. 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.

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/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 PGB, is the responsibility of external contractors. For PISA 2022, the overall management of contractors and implementation was carried out by the Educational Testing Service (ETS) in the United States as part of its responsibility as the **Core A** contractor. The OECD Secretariat worked closely with the International Project Director and Project Manager, to co-ordinate all aspects of implementation. In addition to overall management, Core A was responsible for the computer-delivery platform, instrument development, scaling and analysis, and all data products. As the lead of Core A, ETS worked in co-operation with Westat in the United States for survey operations, cApStAn for translation and verification of the assessment instruments, the International Association for Evaluation of Educational Achievement (IEA) in the Netherlands for the data management software,

The additional tasks related to the implementation of PISA 2022 were carried out by three additional contractors – **Cores B1, B2, B3, C, D, and E**.

The Research Triangle Institute (RTI) in the United States facilitated the development of the mathematics assessment framework as the **Core B1** contractor. ETS also facilitated the development of the background questionnaire frameworks as the **Core B2** contractor. ACT in the United States and Cito in the Netherlands performed the test development for the innovative domain as the **Core B3** contractor. **Core C** focused on sampling and was implemented by Westat in the United States in co-operation with the Australian Council for Educational Research (ACER). **Core D** was managed by cApStAn Linguistic Quality Control in Belgium for linguistic quality control in co-operation with BranTra in Belgium. **Core E** focused on country preparation and implementation support and was managed by the Australian Council for Educational Research (ACER) in Australia.

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

## PISA 2022 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 2022 which are published as *PISA 2022 Results (Volume I): Student performance and Equity in education* (OECD, 2023<sup>[32]</sup>), *PISA 2022 Results (Volume II): Resilient systems, schools and students* (OECD, 2023<sup>[33]</sup>) and subsequent volumes and thematic reports.

Subsequent PISA 2022 result volumes are planned to be published by 2024 as Volumes III, IV, and V on creative thinking, financial literacy, and students' readiness for lifelong learning, respectively.

## References

- Bertling, J. et al. (2020), "A tool to capture learning experiences during COVID-19: The PISA Global Crises Questionnaire Module", *OECD Education Working Papers*, No. 232, OECD Publishing, Paris, <https://doi.org/10.1787/9988df4e-en>. [65]
- Bertling, J. et al. (2020), "A tool to capture learning experiences during COVID-19: The PISA Global Crises Questionnaire Module", *OECD Education Working Papers*, No. 232, OECD Publishing, Paris, <https://doi.org/10.1787/9988df4e-en>. [31]
- OECD (2023), *PISA 2022 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/dfe0bf9c-en>. [64]
- OECD (2023), *PISA 2022 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/dfe0bf9c-en>. [30]
- OECD (2023), *PISA 2022 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris. [32]
- OECD (2023), *PISA 2022 Results (Volume I): Student Performance and Equity in Education*, PISA, OECD Publishing, Paris. [35]
- OECD (2023), *PISA 2022 Results (Volume I): Student Performance and Equity in Education*, PISA, OECD Publishing, Paris. [33]
- OECD (2023), *PISA 2022 Results (Volume II): Resilient Systems, Schools and Students*, PISA, OECD Publishing, Paris. [36]
- OECD (2021), *21st-Century Readers: Developing Literacy Skills in a Digital World*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/a83d84cb-en>. [63]
- OECD (2021), *21st-Century Readers: Developing Literacy Skills in a Digital World*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/a83d84cb-en>. [29]
- OECD (2021), *PISA 2018: Sky's the Limit: Growth Mindset, Students, and Schools in PISA*, PISA, OECD Publishing, Paris, <https://www.oecd.org/pisa/growth-mindset.pdf>. [34]

- OECD (2021), *PISA 2018: Sky's the Limit: Growth Mindset, Students, and Schools in PISA*, PISA, OECD Publishing, Paris, <https://www.oecd.org/pisa/growth-mindset.pdf>. [28]
- OECD (2020), *PISA 2018 Results (Volume IV): Are Students Smart about Money?*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/48ebd1ba-en>. [37]
- OECD (2020), *PISA 2018 Results (Volume IV): Are Students Smart about Money?*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/48ebd1ba-en>. [1]
- OECD (2020), *PISA 2018 Results (Volume V): Effective Policies, Successful Schools*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/ca768d40-en>. [38]
- OECD (2020), *PISA 2018 Results (Volume V): Effective Policies, Successful Schools*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/ca768d40-en>. [2]
- OECD (2020), *PISA 2018 Results (Volume VI): Are Students Ready to Thrive in an Interconnected World?*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/d5f68679-en>. [39]
- OECD (2020), *PISA 2018 Results (Volume VI): Are Students Ready to Thrive in an Interconnected World?*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/d5f68679-en>. [3]
- OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/5f07c754-en>. [40]
- OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/5f07c754-en>. [4]
- OECD (2019), *PISA 2018 Results (Volume II): Where All Students Can Succeed*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/b5fd1b8f-en>. [41]
- OECD (2019), *PISA 2018 Results (Volume II): Where All Students Can Succeed*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/b5fd1b8f-en>. [5]
- OECD (2019), *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/acd78851-en>. [42]
- OECD (2019), *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/acd78851-en>. [6]
- OECD (2017), *PISA 2015 Results (Volume III): Students' Well-Being*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264273856-en>. [43]
- OECD (2017), *PISA 2015 Results (Volume III): Students' Well-Being*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264273856-en>. [9]
- OECD (2017), *PISA 2015 Results (Volume IV): Students' Financial Literacy*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264270282-en>. [44]
- OECD (2017), *PISA 2015 Results (Volume IV): Students' Financial Literacy*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264270282-en>. [8]
- OECD (2017), *PISA 2015 Results (Volume V): Collaborative Problem Solving*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264285521-en>. [56]

- OECD (2017), *PISA 2015 Results (Volume V): Collaborative Problem Solving*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264285521-en>. [7]
- OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264266490-en>. [45]
- OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264266490-en>. [11]
- OECD (2016), *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264267510-en>. [46]
- OECD (2016), *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264267510-en>. [10]
- OECD (2014), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264208070-en>. [48]
- OECD (2014), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264208070-en>. [13]
- OECD (2014), *PISA 2012 Results: Students and Money (Volume VI): Financial Literacy Skills for the 21st Century*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264208094-en>. [49]
- OECD (2014), *PISA 2012 Results: Students and Money (Volume VI): Financial Literacy Skills for the 21st Century*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264208094-en>. [14]
- OECD (2014), *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014): Student Performance in Mathematics, Reading and Science*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264208780-en>. [47]
- OECD (2014), *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014): Student Performance in Mathematics, Reading and Science*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264208780-en>. [12]
- OECD (2013), *PISA 2012 Results: Excellence through Equity (Volume II): Giving Every Student the Chance to Succeed*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264201132-en>. [50]
- OECD (2013), *PISA 2012 Results: Excellence through Equity (Volume II): Giving Every Student the Chance to Succeed*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264201132-en>. [15]
- OECD (2013), *PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264201170-en>. [51]
- OECD (2013), *PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264201170-en>. [16]
- OECD (2013), *PISA 2012 Results: What Makes Schools Successful (Volume IV): Resources, Policies and Practices*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264201156-en>. [52]

- OECD (2011), *PISA 2009 Results: Students On Line: Digital Technologies and Performance (Volume VI)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264112995-en>. [53]
- OECD (2011), *PISA 2009 Results: Students On Line: Digital Technologies and Performance (Volume VI)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264112995-en>. [17]
- OECD (2010), *PISA 2009 Results: Learning to Learn: Student Engagement, Strategies and Practices (Volume III)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264083943-en>. [57]
- OECD (2010), *PISA 2009 Results: Learning to Learn: Student Engagement, Strategies and Practices (Volume III)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264083943-en>. [20]
- OECD (2010), *PISA 2009 Results: Learning Trends: Changes in Student Performance Since 2000 (Volume V)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091580-en>. [59]
- OECD (2010), *PISA 2009 Results: Learning Trends: Changes in Student Performance Since 2000 (Volume V)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091580-en>. [22]
- OECD (2010), *PISA 2009 Results: Overcoming Social Background: Equity in Learning Opportunities and Outcomes (Volume II)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091504-en>. [55]
- OECD (2010), *PISA 2009 Results: Overcoming Social Background: Equity in Learning Opportunities and Outcomes (Volume II)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091504-en>. [19]
- OECD (2010), *PISA 2009 Results: What Makes a School Successful?: Resources, Policies and Practices (Volume IV)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091559-en>. [58]
- OECD (2010), *PISA 2009 Results: What Makes a School Successful?: Resources, Policies and Practices (Volume IV)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091559-en>. [21]
- OECD (2010), *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091450-en>. [54]
- OECD (2010), *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264091450-en>. [18]
- OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264040014-en>. [60]
- OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264040014-en>. [23]
- OECD (2004), *Learning for Tomorrow's World: First Results from PISA 2003*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264006416-en>. [61]

- OECD (2004), *Learning for Tomorrow's World: First Results from PISA 2003*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264006416-en>. [24]
- OECD (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264195905-en>. [62]
- OECD (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264195905-en>. [25]
- OECD/UNESCO Institute for Statistics (2003), *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264102873-en>. [26]
- OECD/UNESCO Institute for Statistics (2003), *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264102873-en>. [66]
- Walker, M. (2011), *PISA 2009 Plus Results: Performance of 15-year-olds in Reading, Mathematics and Science for 10 Additional Participants*, ACER Press, Melbourne. [27]
- walker, M. (2011), *PISA 2009 Plus Result: Performance of 15-years-old-in-Reading, Mathematics and Science for 10 Additional Participants*, ACER Press, Melbourne. [67]




# Annex 1.A. Participation in PISA 2022

Web tables for each chapter can be accessed via the StatLink.

## Annex Table 1.A.1. Chapter 1 Participation in PISA

Table	Title
Web Table 1.A.3	Participating countries and economies in PISA 2022

StatLink  <https://stat.link/2gpth8>

## Annex Table 1.A.2. Participation and main domains in previous PISA cycles

Cycle	Main domain	OECD members / Associates	Partners	Participants
2000	Reading	30	1	31
2000+ (2001/2002)	Reading	3	7	10
2003	Mathematics	33	8	41
2006	Science	39	17	56
2009	Reading	39	25	64
2009+ (2010)	Reading	1	8	9
2012	Mathematics	40	23	63
2015	Science	40	31	71
2018	Reading	40	38	78

Note: Brazil and Thailand have Associate status in PISA.

# 2 The PISA 2022 Integrated Assessment Design

## Introduction

This chapter describes the integrated assessment design for PISA 2022 as well as the processes used by the PISA Core A contractor, Educational Testing Service (ETS) to design the assessment forms for the PISA 2022 cycle.

The cognitive tests for the PISA 2022 cycle included the following:

- a mathematics test, the major domain,
- a reading and a science test, the two minor domains,
- a creative thinking test, the innovative domain, and
- a financial literacy test, an international option.

The development of the mathematics assessment is discussed further in Chapter 3 of this Technical Report. The development of the Creative Thinking domain is presented and discussed in the Chapter 4 of this Technical Report.

## PISA 2022 integrated design

The goals for the integrated assessment design in PISA 2022 included:

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

To meet these goals, the design for PISA 2022 was based on the design and methodological innovations first introduced in the PISA 2015 cycle and the experience with multistage adaptive testing in the PISA 2018 cycle. In contrast to cycles prior to PISA 2015 where scaling was focused on the cycle at hand and required a new scoring transformation each time, the methodology introduced in PISA 2015 incorporated all then available data for scaling and provided a scoring transformation applicable to PISA 2015 as well as future cycles. It provided a more solid basis for linking across cycles and between paper- and computer-based administrations for all cognitive domains and facilitated the development and transition to computer-based adaptive testing.

As a form of adaptive testing particularly well suited for PISA, multistage adaptive testing was adopted in PISA 2018 for the reading literacy domain. This was adopted with the goal to reduce measurement error across heterogeneous populations without overburdening individual respondents. The experience of the 2018 MSAT and taking note of the differences between reading and mathematics allowed further enhancement of the MSAT design for the mathematics CBA assessment in PISA 2022. Taken together, these design and methodological innovations served to improve comparability across countries/economies, improve parameter estimations and the measurement of trends and improve the reliability of inferences made from the data. In addition, as part of the design for PISA 2022, ETS integrated the domain of creative thinking into the assessment design together with the core domains of reading, mathematics, and science.

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

Prior to PISA 2015, the PISA test design focused on keeping the number of students who responded to each item in both the major and minor domains relatively constant. As a result, as shown in Annex Table 2.A.2, the number of items included in the minor domains was significantly lower than the number of items in the major domain (shown in red font for each cycle). Note, for example, that when mathematics was a minor domain in 2000, 2006, and 2009, it contained about 50% of the items used when it was the major domain in 2003, and between 32-44% when it was the major domain in 2012. Furthermore, when reading was a minor domain in 2003 and 2006, it contained only about 20% of the items used when it was the major domain in 2000.

In contrast, under the assessment design for PISA 2022, 197 items were used in the minor domain of reading, which is 80% of the items when reading was last the major domain in 2018 — and there were 115 items in science, which is 63% of the items when it was last the major domain in 2015. Furthermore, the total number of items across the three core domains increased in ten years from 206 in 2012 to 546 in 2022, an increase of 165%.

Altogether, the inclusion of a larger number of items in each minor domain helped to stabilize and improve the measurement of trend by making the construct coverage for each minor domain more comparable to that of a major domain. The target sample size was not increased accordingly, so there was a reduction of the number of student responses per item for the minor domains. However, since trend items are used for minor domains, there typically is sufficient data for each item by combining the information from the current PISA cycle with that from when the subject was a major domain.

Under this approach for measuring trends, each domain goes through a “domain rotation” over four PISA cycles, that begins with a new or revised framework and continues with the two subsequent cycles in which it becomes a minor domain. The rotation concludes, and starts again, with becoming a major domain on the fourth cycle. The end of the full domain cycle involves a revision of the framework to reflect the current thinking about assessment for the new data collection as a major domain. For example, the revised framework for mathematics as the major domain in PISA 2022 and the introduction of computer-based items broadened the construct beyond what was measured in PISA 2012, the last time that mathematics was a major domain. The framework and instruments for mathematics are expected to remain constant for the next two PISA cycles, with the next revision of the mathematics assessment expected for PISA 2033 when mathematics will again be the major domain.

### ***Multistage adaptive testing***

The PISA Governing Board’s (PGB) long-term development strategy for PISA includes the objective of continuing to exploit the advantages of computer-based testing, including the increased use of adaptive testing to further improve measurement accuracy and efficiency, especially at the extremes of the proficiency scale. Additionally, by allowing measurement across a broader range of the ability distribution,

adaptive testing could be viewed as making it possible to better measure a more diverse set of participants, thereby extending the global reach of the PISA assessment.

Multistage adaptive testing (MSAT) was introduced in PISA 2018 for the reading major domain only. In PISA 2022, MSAT was extended to the major domain of mathematics, while a reduced MSAT design was created for the now minor domain of reading. The PISA science assessment does not yet follow an adaptive design and one is foreseen to be implemented in PISA 2025. To prepare the MSAT design for mathematics, during the PISA 2022 field trial, unit order was varied to examine whether the order in which units are presented has any impact on item parameter and proficiency estimation. The results of this study in the field trial showed that unit order did not have a significant impact on item parameters nor on proficiency estimates, supporting the use of an MSAT design for mathematics in the main survey. More information about this aspect is provided under the main survey design section of this chapter.

### **Goals and domain coverage**

The design for the PISA 2022 core assessment was developed to provide participating countries/economies with the following information:

- population proficiency distributions in mathematics, the major domain, that reflect the new PISA 2022 mathematics framework and is linked through trend materials to the framework and scale developed in PISA 2012,
- population proficiency distributions in mathematics process and content subscales,
- population proficiency distributions in the minor domain of reading, linked to the PISA 2018 reading framework through trend items for reading,
- population proficiency distributions in the minor domain of science, linked to the PISA 2015 science framework through trend items for science,
- population proficiency distributions in creative thinking, the innovative domain in PISA 2022,
- correlations among the core domains (mathematics, reading, and science) and the innovative domain (creative thinking),
- correlations between mathematics process and content subscales and the other core domains (reading, and science),
- data to link the two modes of delivery: paper-based and computer-based<sup>1</sup>.

In addition to the three core domains and the innovative domain, the PISA 2022 assessment also included an optional assessment of financial literacy, which was administered only as a computer-based assessment. For countries/economies participating in the optional domain of financial literacy, population distributions linked to the PISA 2018 financial literacy framework through trend items were provided as well as correlations between financial literacy and mathematics and reading domains.

### **Overview of the field trial assessment design**

The PISA 2022 field trial was designed to provide the information needed in preparation for the main survey. Due to the Covid-19 pandemic, many countries/economies had difficulties with either planning, executing, or completing the data collection for the field trial (see the Field Trial section later in this chapter).

As with the PISA 2018 field trial, the PISA 2022 field trial was designed to verify trend and new items and the feasibility of the integrated design planned for the main survey. In particular, it was designed to verify the feasibility of the new MSAT design for mathematics planned for the main survey and the reduced MSAT design for reading. To ensure appropriate sampling of content, scaling of items and, improved adaptation to student proficiency, the PISA MSAT design offers many alternative options for the selection and delivery

of many pre-assembled testlets (i.e. a set of items containing several units) of varying difficulty. As part of the design, units need to be assigned to more than one testlet in different test positions. Thus, while the order of items within a unit does not change, the position of a unit across testlets can be different. For example, a certain unit can be presented as the first unit in some testlets, but as the second unit in others. Therefore, it is important to verify that the psychometric properties of the items and units are invariant when used in different positions (i.e. absence of item/unit position effects). Furthermore, the same unit can be surrounded by different units in different testlets across stages of the MSAT, so that testlets of different difficulty levels are created while ensuring links between them.

The observation of order effects in early PISA cycles (prior to 2015) had led to the assumption that intact cluster positions were needed for parameter invariance to hold. However, a rescaling study conducted on the joint database of all historical PISA data collected between 2000 and 2012 showed good stability of item parameters overall across multiple survey cycles even though over time there were deviations from the strict application of the “intact cluster” paradigm (von Davier et al., 2019<sup>[1]</sup>). The PISA 2022 field trial was designed to provide additional information regarding item parameter invariance under variable unit positions. To that effect, the field trial collected data to study unit order effects by manipulating fixed and variable positions within 30-minute (intact) clusters, and students were randomly assigned to three groups with different unit orders.

For the PISA 2022 field trial, a unit was again considered to represent the minimum granular size of item sets at which adaptiveness can take place. Units consist of a set of items based on a common stimulus or stimuli that can be considered as the organizing grain size that can be assigned randomly or guided by adaptiveness. Although within-unit adaptiveness would be possible in principle, no variations were introduced within a unit. However, the sequence of units within a cluster can be changed to examine parameter invariance relative to unit position. Examining and ensuring parameter invariance at the unit level was a necessary condition for the PISA 2022 mathematics assessment to be delivered in adaptive mode.

With this in mind, the goals of the field trial design included:

- evaluation of the invariance of item parameters compared to previous PISA cycles (both CBA and PBA),
- evaluation of the invariance of item parameters regarding the positions of intact units; that is, a comparison of stability of item parameters between 30-minute clusters found in prior PISA cycles versus varying positions of smaller collections of units to examine the feasibility of introducing MSAT for mathematics in the main survey,
- obtaining preliminary item parameters for the evaluation of new mathematics, financial literacy, and creative thinking items, and for the selection of a final set of items used in the main survey for these new units,
- evaluating sampling and survey operations,
- assessing how well the computer platform functions within and across participating countries/economies.

Like the main survey design, the field trial design for PISA 2022 implemented one CBA design including mathematics, reading, science as core domains, creative thinking as innovative domain, and financial literacy as the optional domain. In addition, the field trial design also included two PBA designs that involved the three core domains of mathematics, reading, and science. One PBA design was the same as implemented in PISA 2018. The other, new PBA design was developed for newly participating countries/economies. The new PBA instrument was the same one that was used for PISA for Development<sup>2</sup>.

The standard design for countries/economies choosing computer delivery for the assessment was to select a minimum of 28 schools for the field trial and select 71-72 students within each school. This design

resulted in a sample size of approximately 2,000 assessed students. Alternative designs to achieve the same sample size were available for participants having difficulty in finding enough large schools where to implement this design.

Countries/economies that chose to participate using only paper-and-pencil forms had a reduced sample-size requirement. The goals for these participants were mainly focused on testing operations and data-processing related procedures. For both the PBA and new PBA designs, these participants selected 25 schools with 36 students from each school for a total field trial sample of approximately 900 assessed students.

### ***Field trial CBA design***

The computer-based assessment (CBA) design for the field trial organized the items into 69 different test forms and students into three groups. Students in groups 1 and 3 took fixed-unit order (FUO) forms, while students in group 2 took variable-unit order (VUO) forms. The standard field trial CBA design is shown in Table 2.1. Each test form consisted of at most two domains, resulting in at least one hour of assessment time per domain, with a total of two hours of testing time per student. Each cluster consisted of multiple units, and the ordering of the units was always fixed and consistent in FUO forms. In contrast, ordering of the units was varied across VUO forms. For example, cluster M1 cluster in form 19 had a different ordering of units compared to the ordering of units in cluster M1 in form 25. More specifically, students in group 1 took forms 1–18 with trend items in mathematics, reading, and science. Group 2 took 24 forms (forms 19–42) with both new and trend mathematics items. Group 3 took 27 forms with either only new mathematics items (forms 43–54) or new mathematics and creative thinking items (forms 55–69). Students in group 1 who took reading were administered the reduced MSAT design discussed later in this chapter. Furthermore, the same set of 65 sentences from the 2018 Main Survey were used to measure reading fluency as part of the Reading scale.

### ***Field trial PBA designs***

As noted, there were two PBA instruments offered this PISA cycle. The first PBA design was a version administered by only one participant and contained the same trend clusters that were administered in PISA 2015 and PISA 2018 for paper-based participants. The second PBA was new for this PISA cycle. However, the materials have previously been administered in PISA for Development and were successfully linked to the PISA scales as there are items common to both instruments. This new PBA instrument was administered by all other PBA participants. Under the first PBA design, students were randomly assigned one of the 18 PBA forms that contained trend items from two of the three core domains for PISA – reading, mathematics, and science. This design is shown in Table 2.2.

Students in countries/economies that chose the second, new PBA design were randomly assigned one of 12 new PBA forms that contained trend items from two of the three core domains for PISA – mathematics, science, and reading/reading fluency. This design is also shown in Overview of the main survey assessment design in Table 2.2.

The assessment design for PISA 2022 was planned so that the total testing time was two hours for each student, followed by a student background questionnaire. An overview of the flow of the integrated design for the PISA 2022 main survey is presented in Figure 2.1.

Table 2.1. Field trial computer-based assessment design

	Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
GROUP 1 CBA Trend FUO (Forms 01-18)	1	S1	S4	M1	M2
	2	S3	S6	M3	M4
	3	S5	S2	M5	M6ab
	4	M2	M3	S2	S5
	5	M4	M5	S4	S1
	6	M6ab	M1	S6	S3
	7	M1	M4	R adaptive	R adaptive
	8	M3	M6ab	R adaptive	R adaptive
	9	M5	M2	R adaptive	R adaptive
	10	R adaptive	R adaptive	M2	M5
	11	R adaptive	R adaptive	M4	M1
	12	R adaptive	R adaptive	M6ab	M3
	13	R adaptive	R adaptive	S1	S2
	14	R adaptive	R adaptive	S3	S4
	15	R adaptive	R adaptive	S5	S6
	16	S2	S3	R adaptive	R adaptive
	17	S4	S5	R adaptive	R adaptive
	18	S6	S1	R adaptive	R adaptive
GROUP 2 CBA Trend M/New M VUO (Forms 19-42)	19	M1	M14	M12	M7
	20	M2	M16	M14	M9
	21	M3	M18	M16	M11
	22	M4	M8	M18	M13
	23	M5	M10	M8	M15
	24	M6ab	M12	M10	M17
	25	M13	M1	M10	M9
	26	M15	M2	M12	M11
	27	M17	M3	M14	M13
	28	M7	M4	M16	M15
	29	M9	M5	M18	M17
	30	M11	M6ab	M8	M7
	31	M11	M18	M1	M8
	32	M13	M8	M2	M10
	33	M15	M10	M3	M12
	34	M17	M12	M4	M14
	35	M7	M14	M5	M16
	36	M9	M16	M6ab	M18
	37	M16	M17	M15	M1
	38	M18	M7	M17	M2
	39	M8	M9	M7	M3
	40	M10	M11	M9	M4
	41	M12	M13	M11	M5
	42	M14	M15	M13	M6ab
GROUP 3 CBA New M/CRT FUO (Forms 43-69)	43	M7	M8	M10	M14
	44	M8	M9	M11	M15
	45	M9	M10	M12	M16
	46	M10	M11	M13	M17
	47	M11	M12	M14	M18
	48	M12	M13	M15	M7
	49	M13	M14	M16	M8
	50	M14	M15	M17	M9
	51	M15	M16	M18	M10
	52	M16	M17	M7	M11
	53	M17	M18	M8	M12
	54	M18	M7	M9	M13
	55	M7	M13	CT1	CT2
	56	M8	M14	CT2	CT3
	57	M9	M15	CT3	CT4
	58	M10	M16	CT4	CT5
	59	M11	M17	CT5	CT1
	60	CT3	CT5	M14	M9
	61	CT4	CT1	M15	M10
	62	CT5	CT2	M16	M11
	63	CT1	CT3	M17	M12
	64	CT2	CT4	M18	M7
	65	CT1	CT2	CT3	CT5
	66	CT2	CT3	CT4	CT1
	67	CT3	CT4	CT5	CT2
	68	CT4	CT5	CT1	CT3
	69	CT5	CT1	CT2	CT4

FUO = fixed unit order; VUO = variable unit order

Where: R adaptive represents CBA trend reading units (containing trend and new items from 2018)

M7-M18 represent CBA new mathematics clusters

M1-M6ab represent CBA trend mathematics clusters (in the 2022 FT, all CBA participants administered both M6a and M6b)

S1-S6 represent CBA trend science clusters (containing trend and new items from 2015)

CT1-CT5 represent CBA new creative thinking clusters

Table 2.2. Field trial paper-based assessment designs

Design 1 - PBA Design							
		Booklets	Cluster 1	Cluster 2	Cluster 3	Cluster 4	
900 assessed students (25 schools, 36 students per school)	P=0.47	1	PS1	PS4	PM1	PM2	
		2	PS3	PS6	PM3	PM4	
		3	PS5	PS2	PM5	PM6b	
		4	PM2	PM3	PS2	PS5	
		5	PM4	PM5	PS4	PS1	
		6	PM6b	PM1	PS6	PS3	
	P=0.47	7	PM1	PM4	PR1	PR2	
		8	PM3	PM6b	PR3	PR4	
		9	PM5	PM2	PR5	PR6b	
		10	PR2	PR3	PM2	PM5	
		11	PR4	PR5	PM4	PM1	
		12	PR6b	PR1	PM6b	PM3	
	P=0.06	13	PR1	PR4	PS1	PS2	
		14	PR3	PR6b	PS3	PS4	
		15	PR5	PR2	PS5	PS6	
		16	PS2	PS3	PR2	PR5	
		17	PS4	PS5	PR4	PR1	
		18	PS6	PS1	PR6b	PR3	

Design 2 - "new" PBA design							
		Booklets	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
900 assessed students (25 schools, 36 students per school)	P=1.00	1	RC1	R1	R2	S1	S2
		2	S2	S3	RC2	R2	R3
		3	RC3	R3	R4	S3	S4
		4	S4	S1	RC4	R4	R1
		5	S1	S2	M1	M2	
		6	M2	M3	S2	S3	
		7	S3	S4	M3	M4	
		8	M4	M1	S4	S1	
		9	M1	M2	RC1	R1	R2
		10	RC2	R2	R3	M2	M3
		11	M3	M4	RC3	R3	R4
		12	RC4	R4	R1	M4	M1

Notes: Design 1: where:

PR1-PR6b represent PBA trend reading clusters (the participant only administered R6b) - same clusters from 2015 and 2018

PM1-PM6b represent PBA trend mathematics clusters (the participant only administered M6b) - same clusters from 2015 and 2018

PS1-PS6 represent PBA trend science clusters (same clusters from 2015 and 2018)

Design 2: where:RC1-RC4 represent reading components clusters

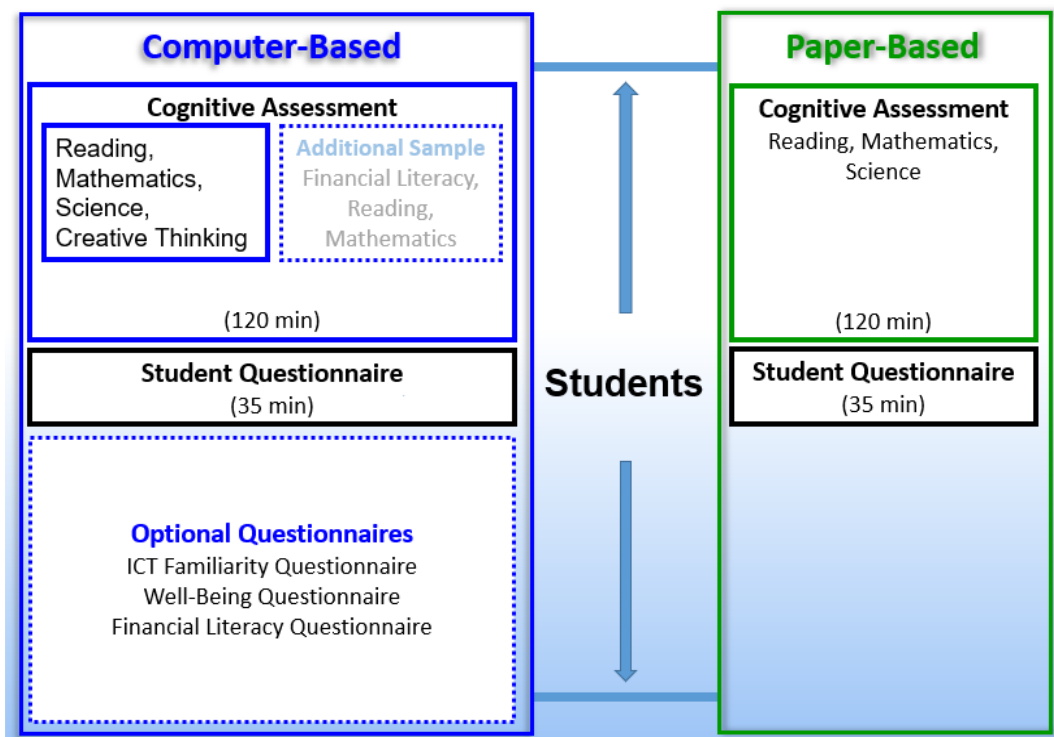
R1-R4 represent new PBA reading clusters

M1-M4 represent new PBA mathematics clusters

S1-S4 represent new PBA science clusters Booklets 5-8 did not contain a reading components cluster



Figure 2.1. Overview of the PISA 2022 main survey integrated design



### ***Paper-based integrated designs***

For the participant in the first PBA design, the main survey included the same 18 forms as in the field trial assessment design, but sample size requirements differed. The main survey PBA design is shown in Table 2.3. The PBA test forms did not include any newly developed items. Each form included two of the three core domains with two 30-minute clusters for each domain assessed. As a result, all students were administered four clusters, 47% of participating students were administered two clusters of science items and two clusters of mathematics items, 47% were administered two clusters of mathematics and two clusters of reading, and 6% were administered two clusters of reading and two clusters of science. The PBA was to be administered to 35 students in each of 150 schools, resulting in a total sample size of 5,250 assessed students.

The main survey assessment design for countries/economies that chose the new PBA design included 12 forms (see Table 2.3.) and was the same as for the field trial. These PBA test forms consisted of existing items from PISA for Development. Each form included two of the three core domains with two 30-minute clusters for each domain assessed. Students were administered a randomly selected form. As a result, 33% of participating students were administered two clusters of reading items and two clusters of science items, 33% were administered two clusters of science and two clusters of mathematics, and 33% were administered two clusters of mathematics and two clusters of reading. As with the first PBA design, the new PBA design was to be administered to 35 students in each of 150 schools, resulting in a total of 5,250 assessed students.

**Table 2.3. Main survey paper-based assessment designs**

The field trial and main survey paper-based assessment designs were the same with respect to the items/units and clusters, number of booklets, and the order of the clusters within the booklets.

Design 1 - PBA Design							
		Booklets	Cluster 1	Cluster 2	Cluster 3	Cluster 4	
5,250 assessed students (150 schools, 35 students per school)	P=0.47	1	PS1	PS4	PM1	PM2	
		2	PS3	PS6	PM3	PM4	
		3	PS5	PS2	PM5	PM6b	
		4	PM2	PM3	PS2	PS5	
		5	PM4	PM5	PS4	PS1	
		6	PM6b	PM1	PS6	PS3	
	P=0.47	7	PM1	PM4	PR1	PR2	
		8	PM3	PM6b	PR3	PR4	
		9	PM5	PM2	PR5	PR6b	
		10	PR2	PR3	PM2	PM5	
		11	PR4	PR5	PM4	PM1	
		12	PR6b	PR1	PM6b	PM3	
	P=0.06	13	PR1	PR4	PS1	PS2	
		14	PR3	PR6b	PS3	PS4	
		15	PR5	PR2	PS5	PS6	
		16	PS2	PS3	PR2	PR5	
		17	PS4	PS5	PR4	PR1	
		18	PS6	PS1	PR6b	PR3	

Design 2 - new PBA Design							
		Booklets	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
5,250 assessed students (150 schools, 35 students per school)	P=0.33	1	RC1	R1	R2	S1	S2
		2	S2	S3	RC2	R2	R3
		3	RC3	R3	R4	S3	S4
		4	S4	S1	RC4	R4	R1
	P=0.33	5	S1	S2	M1	M2	
		6	M2	M3	S2	S3	
		7	S3	S4	M3	M4	
		8	M4	M1	S4	S1	
	P=0.33	9	M1	M2	RC1	R1	R2
		10	RC2	R2	R3	M2	M3
		11	M3	M4	RC3	R3	R4
		12	RC4	R4	R1	M4	M1

Where: Design 1: PR1-PR6b represent PBA trend reading clusters (the participant only administered R6b) - same clusters from 2015 and 2018  
 PM1-PM6b represent PBA trend mathematics clusters (the participant only administered M6b) - same clusters from 2015 and 2018  
 PS1-PS6 represent PBA trend science clusters (same clusters from 2015 and 2018)

Where: Design 2: RC1-RC4 represent reading components clusters  
 R1-R4 represent new PBA reading clusters  
 M1-M4 represent new PBA mathematics clusters  
 S1-S4 represent new PBA science clusters  
 Booklets 5-8 did not contain a reading components cluster

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

For CBA participants that also administered the creative thinking assessment, the main survey included 66 forms (forms 01-66) which are shown in Table 2.4. . Under the full integrated design that included all four domains, 94% of the sampled students responded to 60 minutes of mathematics items, 39% responded to reading items, 39% to science items, and 28% responded to creative thinking items. As in PISA 2018, sixty-five reading fluency items assigned to six blocks were used. Each student taking reading received two blocks of sentences which were rotated as shown in Table 2.4. .

For countries/economies not participating in the creative thinking assessment, only 36 forms were included in the design (forms 01-36). The percentages for this alternative design are also represented in Figure 2.2.

#### *Main survey multistage adaptive testing design: Mathematics and Reading*

The MSAT design that was implemented for mathematics in the PISA 2022 main survey was built upon the MSAT design used for reading in PISA 2018. However, using the experience from PISA 2018 and the differences between mathematics and reading, it was possible to enhance the following four areas:

1. *Balancing the MSAT design.* A fully balanced design was implemented so that each item occurred in every stage, this to further address potential position effects. This feature is similar to the balanced incomplete block (BIB) designs used in previous, non-adaptive PISA cycles.
2. *More adaptivity.* A third level of difficulty was introduced in the third stage, which was possible because there were more machine-scored items and smaller units in mathematics than there were in reading.
3. *Linear component.* A hybrid design with an adaptive and linear component was used so that the probability layer used in the PISA 2018 MSAT design for reading could be eliminated. The probability layer used determined the difficulty of the next set of items to be administered, with a low probability assigned to a misrouting. Instead of this probability layer, 25% of students were administered a linear test to avoid the intentional misrouting of students to items that would be either too easy or too difficult for them).
4. *Automated assembly.* Formal methods for optimal design and test assembly were employed by making use of linear programming techniques, which provided a principled approach to support the decision-making process for the MSAT design.

Since reading was not the major domain this cycle, the MSAT reading design used for PISA 2022 was a reduced version of the MSAT design used in PISA 2018. That is, the same number of stages and adaptive levels were used, but with a smaller item pool (about 25% fewer items, 196 instead of 245 items) and fewer testlets (30 instead of 40 testlets). As in PISA 2018, each student assessed in reading received seven units. In design A (75%), students take 2, 3, and 2 reading units across the three stages from three sets of units, whereas students take 2, 2, and 3 reading units, respectively, in design B (25%) where the unit sets for the last two stages are reversed compared to design A. The same probability layer was used as in PISA 2018 for routing students through different MSAT paths (see PISA 2018 Tech Report, Chapter 2). In PISA 2022, each student assessed in reading responded to 35-42 reading items, while in PISA 2018 the range was 33-40 items. The PISA 2022 design still allowed students to take approximately the same number of items within the same amount of assessment time.

Table 2.4. Main survey computer-based assessment design

Percentage of Students	Forms	Fluency	Cluster 1	Cluster 2	Fluency	Cluster 3	Cluster 4
35% (No CT= 48%)	1		M <sub>(adaptive)</sub>		fi1	R <sub>(adaptive)</sub>	
	2		M <sub>(adaptive)</sub>		fi2	R <sub>(adaptive)</sub>	
	3		M <sub>(adaptive)</sub>		fi3	R <sub>(adaptive)</sub>	
	4		M <sub>(adaptive)</sub>		fi4	R <sub>(adaptive)</sub>	
	5		M <sub>(adaptive)</sub>		fi5	R <sub>(adaptive)</sub>	
	6		M <sub>(adaptive)</sub>		fi6	R <sub>(adaptive)</sub>	
	7	fi7		R <sub>(adaptive)</sub>			M <sub>(adaptive)</sub>
	8	fi8		R <sub>(adaptive)</sub>			M <sub>(adaptive)</sub>
	9	fi9		R <sub>(adaptive)</sub>			M <sub>(adaptive)</sub>
	10	fi10		R <sub>(adaptive)</sub>			M <sub>(adaptive)</sub>
	11	fi11		R <sub>(adaptive)</sub>			M <sub>(adaptive)</sub>
	12	fi12		R <sub>(adaptive)</sub>			M <sub>(adaptive)</sub>
35% (No CT= 48%)	13		M <sub>(adaptive)</sub>			S1	S2
	14		M <sub>(adaptive)</sub>			S2	S3
	15		M <sub>(adaptive)</sub>			S3	S4
	16		M <sub>(adaptive)</sub>			S4	S5
	17		M <sub>(adaptive)</sub>			S5	S6
	18		M <sub>(adaptive)</sub>			S6	S1
	19		S1	S3			M <sub>(adaptive)</sub>
	20		S2	S4			M <sub>(adaptive)</sub>
	21		S3	S5			M <sub>(adaptive)</sub>
	22		S4	S6			M <sub>(adaptive)</sub>
	23		S5	S1			M <sub>(adaptive)</sub>
	24		S6	S2			M <sub>(adaptive)</sub>
2% (No CT= 4%)	25	fi1	R <sub>(adaptive)</sub>			S1	S2
	26	fi2	R <sub>(adaptive)</sub>			S2	S3
	27	fi3	R <sub>(adaptive)</sub>			S3	S4
	28	fi4	R <sub>(adaptive)</sub>			S4	S5
	29	fi5	R <sub>(adaptive)</sub>			S5	S6
	30	fi6	R <sub>(adaptive)</sub>			S6	S1
	31		S1	S3	fi7		R <sub>(adaptive)</sub>
	32		S2	S4	fi8		R <sub>(adaptive)</sub>
	33		S3	S5	fi9		R <sub>(adaptive)</sub>
	34		S4	S6	fi10		R <sub>(adaptive)</sub>
	35		S5	S1	fi11		R <sub>(adaptive)</sub>
	36		S6	S2	fi12		R <sub>(adaptive)</sub>
24% (No CT= NA)	37		M <sub>(adaptive)</sub>			CT1	CT2
	38		M <sub>(adaptive)</sub>			CT2	CT3
	39		M <sub>(adaptive)</sub>			CT3	CT4
	40		M <sub>(adaptive)</sub>			CT4	CT5
	41		M <sub>(adaptive)</sub>			CT5	CT1
	42		CT2	CT4			M <sub>(adaptive)</sub>
	43		CT3	CT5			M <sub>(adaptive)</sub>
	44		CT4	CT1			M <sub>(adaptive)</sub>
	45		CT5	CT2			M <sub>(adaptive)</sub>
	46		CT1	CT3			M <sub>(adaptive)</sub>
2% (No CT= NA)	47	fi1	R <sub>(adaptive)</sub>			CT1	CT2
	48	fi2	R <sub>(adaptive)</sub>			CT2	CT3
	49	fi3	R <sub>(adaptive)</sub>			CT3	CT4
	50	fi4	R <sub>(adaptive)</sub>			CT4	CT5
	51	fi5	R <sub>(adaptive)</sub>			CT5	CT1
	52		CT2	CT4	fi7		R <sub>(adaptive)</sub>
	53		CT3	CT5	fi8		R <sub>(adaptive)</sub>
	54		CT4	CT1	fi9		R <sub>(adaptive)</sub>
	55		CT5	CT2	fi10		R <sub>(adaptive)</sub>
	56		CT1	CT3	fi11		R <sub>(adaptive)</sub>
2% (No CT= NA)	57		S1	S3		CT1	CT2
	58		S2	S4		CT2	CT3
	59		S3	S5		CT3	CT4
	60		S4	S6		CT4	CT5
	61		S5	S1		CT5	CT1
	62		CT2	CT4		S1	S2
	63		CT3	CT5		S2	S3
	64		CT4	CT1		S3	S4
	65		CT5	CT2		S4	S5
	66		CT1	CT3		S5	S1

Where: R(adaptive) represents the computer-based reading assessment (trend) in an adaptive design  
M(adaptive) represents the computer-based mathematics assessment (trend and new) in an adaptive design  
S1-S6 represent the computer-based science clusters (trend)  
CT1-CT5 represent the computer-based creative thinking clusters (new)  
fi1-fi12 represent the computer-based reading fluency clusters (trend and new items)

**Figure 2.2. Overview of main survey computer-based MSAT design - with creative thinking and without creative thinking**

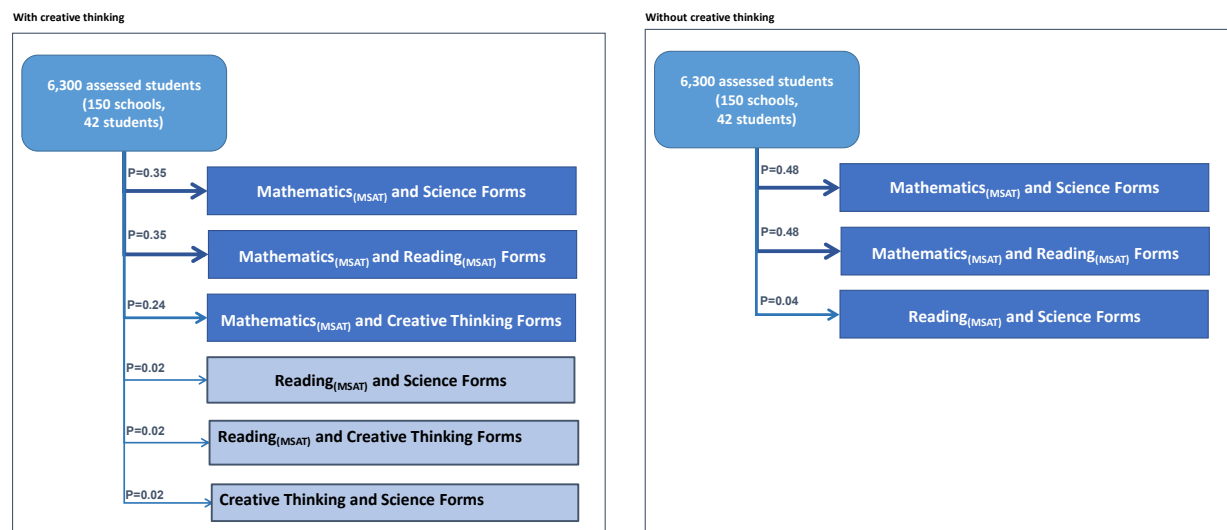


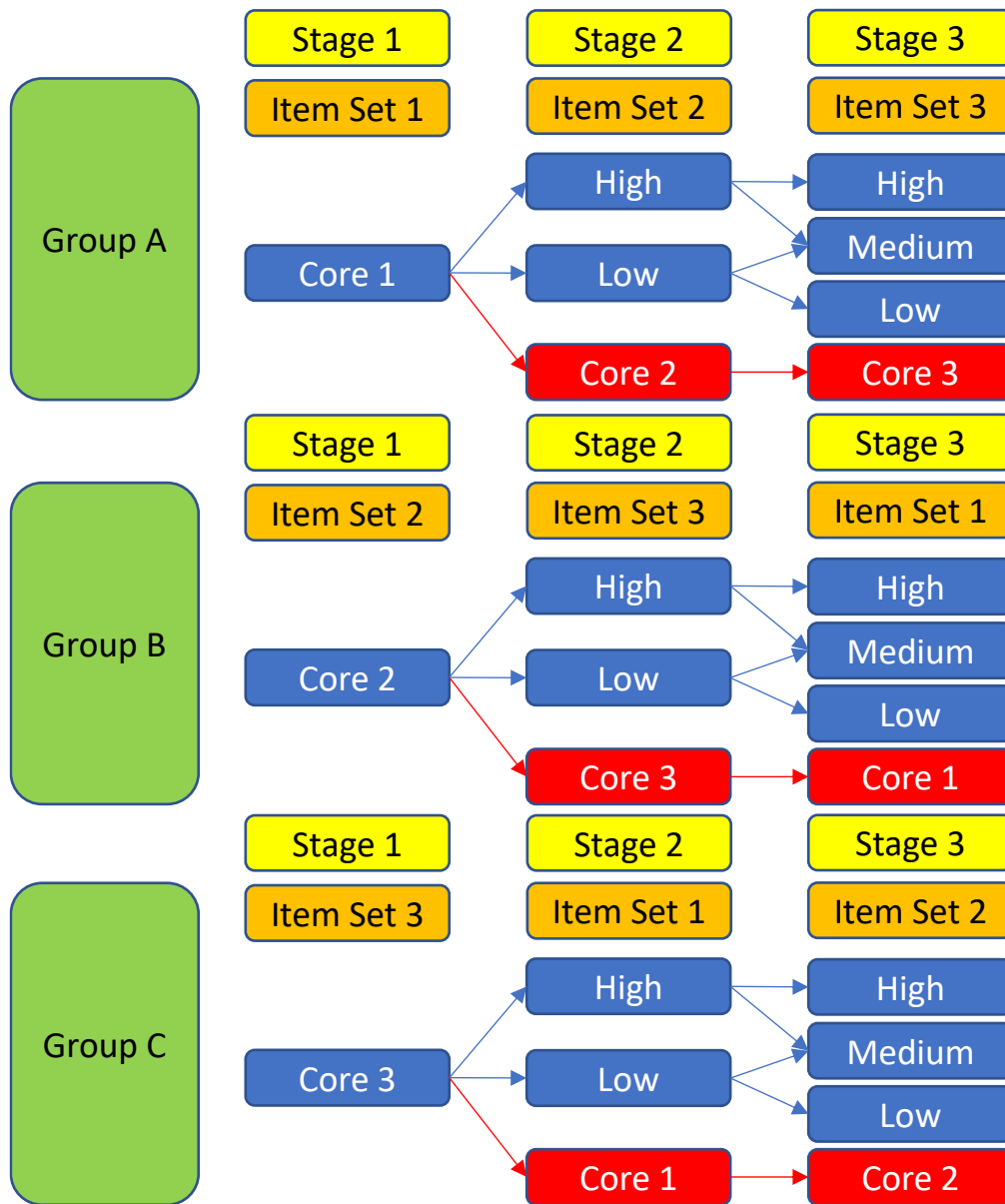
Figure 2.3 shows an overview of the hybrid MSAT design used for mathematics in the PISA 2022 main survey. The MSAT design for mathematics consisted of three stages and 234 mathematics items from a total of 99 units. The items were divided into three equivalent and mutually exclusive item sets, each consisting of 78 items from 33 units. From each item set, 16 testlets of nine or 10 items were created within each stage, so across the three item sets and three stages, there was a total  $16 \times 3 \times 3 = 144$  testlets. Each student took one testlet in each stage, so the total number of mathematics items taken by each student ranged from 28-30. Within-stage linking was accomplished by having each item appear two, or sometimes three, times across testlets associated with each stage and each group (but no more than seven times overall). For students taking the adaptive part of the design, stage 1 consisted of a core testlet of medium difficulty, stage 2 consisted of high- or low-difficulty testlets, and stage 3 consisted of high-, medium-, or low-difficulty testlets administered in a rotating order to constitute three sets of equivalent instruments that were assigned to three groups of randomly selected students (A, B, and C). For students that were assigned to the linear part of the design, after the stage 1 core testlet, they proceeded to take a core testlet from the other item sets at each subsequent stage. Figure 2.4 shows the testlet structure for one group (Group A) and the item set associated with that group, as well as four example paths that a student could take under the adaptive part of the design.

The total number of paths in the hybrid MSAT design for mathematics was 240 (see Annex Table 2.A.3). For the adaptive component, there were 192 total paths since every testlet in stage 1 was associated with four possible paths (going from Stage 1 > Stage 2 > Stage 3):

1. Core > Low > Low
2. Core > Low > Medium
3. Core > High > Medium
4. Core > High > High

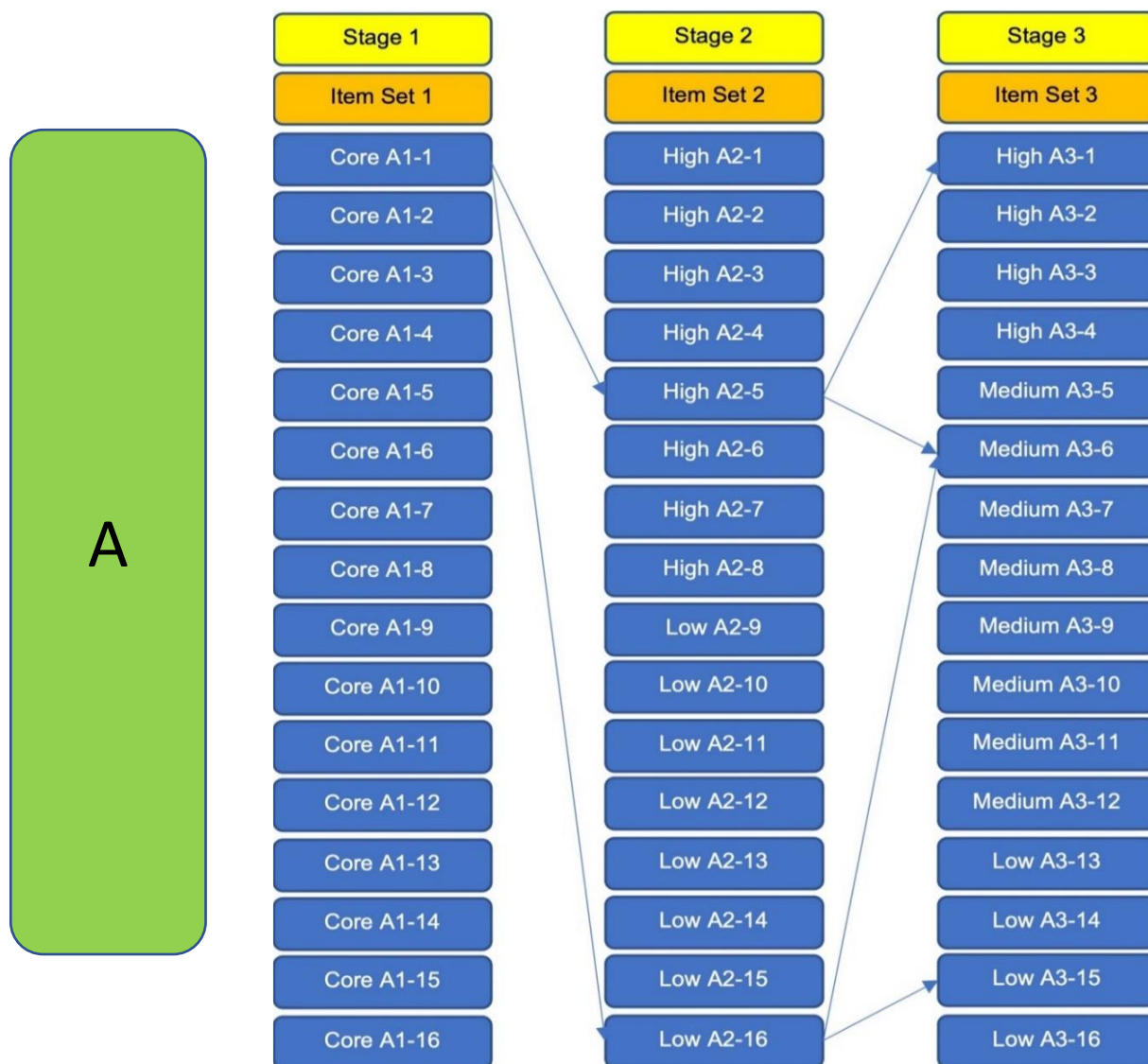
For the linear component, a simplified design was chosen where each testlet in stage 1 was associated with one fixed path that resulted in 48 linear forms. The forms are shown in Annex Table 2.A.4.

Figure 2.3. Overview of the hybrid main survey computer-based MSAT design for mathematics



Where:  
 Groups A, B, and C represent groups of randomly selected students  
 Blue represents adaptive parts - taken by 75% of students  
 Red represents linear parts - taken by 25% of students

Figure 2.4. Example testlet structure across stages for one group



Where:

A represents one group of randomly selected students. The structure is the same for groups B and C, and for the item sets associated with each stage for those groups.

The arrows represent an example of four possible paths. By design, some combinations of testlets were not allowed.

The difficulty of the testlets was targeted by using subsets of the item pool as the statistical target. The average difficulty in stage 1 was targeted by using 100% of the items. At stage 2, low difficulty testlets were targeted by using 75% of the easiest items, and high difficulty levels were targeted by using 75% of the hardest items. At stage 3, a similar approach was taken for low, medium, and high difficulty levels by using 50% of the easiest items, 50% of medium difficulty items, and 50% of the hardest items.

Technically, this targeting was accomplished by using the test information function (TIF) of the relevant subsets of items as the statistical target in the assembly. However, since differences in difficulty can still arise when only the TIF is used [see e.g. Ali and van Rijn (2016<sub>[2]</sub>)], constraints on the test characteristic curve (TCC) were used as well. The method resulted in the high difficulty testlets at stage 3 being more difficult than the high difficulty testlets at stage 2, and the low difficulty testlets at stage 3 were less difficult

than the low difficulty testlets at stage 2, which is ideal because more is known about a student's mathematical proficiency after two stages of assessment.

Additionally, to avoid students experiencing a large shift in difficulty levels between stages, as well as to keep the number of possible paths to a more reasonable number, students who received a low difficulty testlet in stage 2 could not be routed to a high difficulty testlet in stage 3, and students who received a high difficulty testlet in stage 2 could not be routed to a low difficulty testlet in stage 3. The effect of restricting the possible paths is minimal because there is a considerable amount of overlap in the difficulty ranges of testlets of adjacent difficulty (i.e. low/medium and medium/high).

Cut-off values for determining how to route students were identified by first computing the intersections of the average information functions of the testlets. On the PISA mathematics scale, the intersection of low and high difficulty testlets at stage 2 was found to be 495. At stage 3, the intersection between low and medium was found to be 425, and between medium and high was 550. Once these values were identified, the inverse TCC was used to determine the cut scores based on the items within each testlet that could be automatically scored. The cut scores were used to determine a student's path as each stage was completed. Simulation studies showed that this approach would result in about one third of students being routed to each of the difficulty levels at stage 3 for a country/economy that performs around 500 – the midpoint of the scale.

Annex Table 2.B.1 of this chapter provides a list of the cut scores, including the maximum score from machine-coded items and the maximum possible score, for each core testlet. Annex Table 2.C.1 of this chapter shows cut scores for each adaptive path, including the maximum score from machine-coded items and the maximum possible score. These cut scores are based on the number of raw score points obtained on the machine scored items alone.

### ***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 trend items (i.e. items developed prior to PISA 2015) in each core domain and had a reduced reading load. The UH form contained about half as many items as the other forms, with each cluster including from seven to nine items. In PISA 2022 the UH form was comprised of about 53% mathematics, 21% reading, and 26% science items.

The UH form included two 15-minute clusters of mathematics (MU1 and MU2), one 15-minute cluster of reading (RU1) and one 15-minute cluster of science (SU1). The assignment of this form followed the approach described previously for the assignment of the base test form. The UH form was assigned base form 99 (as shown in Table 2.5.).

**Table 2.5. Main survey computer-based UH form design**

Form	Cluster 1	Cluster 2	Cluster 3	Cluster 4
99 (UH)	MU1	MU2	RU1	SU1

Where M = mathematics, R = reading, and S = science

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



## Assessment of financial literacy

The assessment of financial literacy was again offered as an international option in PISA 2022. The cognitive instruments included trend items from the PISA 2012, PISA 2015, and PISA 2018 assessments, plus a few new units that were developed for PISA 2022. Financial literacy was administered only as a computer-based assessment to an additional sample of students at the same schools sampled for PISA.

As in PISA 2018, the financial literacy assessment was administered to a separate sample of PISA-eligible students who took, in addition to the financial literacy assessment, a combination of reading or mathematics items. The total testing time for each student was two hours (120 minutes). The sample of students who took the financial literacy assessment are referred to as the “Financial Literacy sample”.

### Field trial design for the financial literacy assessment

For the 2022 field trial of the financial literacy assessment, the main sample was augmented by adding a sample of approximately 253 students who were assigned one of the 12 financial literacy testing forms. These forms included 60 minutes of financial literacy items and either 60 minutes of reading items or 60 minutes of mathematics items. These were based on using two financial literacy clusters (F1 and F2), MSAT reading items, and six of the seven trend mathematics clusters (M1 to M6ab). The design is shown in Table 2.6. . The 12 financial literacy forms were administered to Group 1 students (FUO) and each form was administered to about 32 students within each country/economy.

**Table 2.6. Field trial computer-based financial literacy design**

Form		Cluster 1	Cluster 2		Cluster 3	Cluster 4
70		M1	M2		F1	F2
71		M3	M4		F2	F1
72		M5	M6ab		F1	F2
73	fl1	R <sub>(adaptive)</sub>			F2	F1
74	fl2	R <sub>(adaptive)</sub>			F1	F2
75	fl3	R <sub>(adaptive)</sub>			F2	F1
76		F2	F1		M2	M5
77		F1	F2		M4	M1
78		F2	F1		M6ab	M3
79		F1	F2	fl4	R <sub>(adaptive)</sub>	
80		F2	F1	fl5	R <sub>(adaptive)</sub>	
81		F1	F2	fl6	R <sub>(adaptive)</sub>	

Where:

F1-F2 represent the computer-based financial literacy clusters (new and trend)

R<sub>(adaptive)</sub> represents the computer-based reading assessment (trend and new) in an adaptive design

M1-M6ab represent the computer-based mathematics trend clusters

fl1-fl6 represent reading fluency clusters

### Main survey financial literacy design

For the main survey, countries/economies participating in the financial literacy assessment were required to assess 1,650 additional students. Each student that took the financial literacy assessment took 60 minutes of financial literacy items, and then either mathematics or reading items. Students taking the financial literacy assessment did not take any of the science items and therefore they do not have science literacy proficiency estimates.

The main survey version of the assessment instruments included 46 financial literacy items, of which 41 were trend items and 5 were new items. These items were organized into two 30-minute clusters of financial literacy (F1 and F2) that were rotated into eight forms each containing 60 minutes of financial literacy and 60 minutes of either MSAT mathematics or MSAT reading items, as shown in Table 2.7. .

**Table 2.7. Main survey computer-based financial literacy design**

Form	Cluster 1	Cluster 2	Fluency	Cluster 3	Cluster 4
67	$M_{(adaptive)}$			F1	F2
68	$M_{(adaptive)}$			F2	F1
69	F1	F2		$M_{(adaptive)}$	
70	F2	F1		$M_{(adaptive)}$	
71	$R_{(adaptive)}$			F1	F2
72	$R_{(adaptive)}$			F2	F1
73	F1	F2	fl7	$R_{(adaptive)}$	
74	F2	F1	fl8	$R_{(adaptive)}$	

Where:

F1-F2 represent the computer-based financial literacy clusters (new and trend)

$R_{(adaptive)}$  represents the computer-based reading assessment (trend and new) in an adaptive design

$M_{(adaptive)}$  represents the computer-based mathematics assessment (trend and new) in an adaptive design

fl7-fl8 represent reading fluency clusters

### Assigning mathematics units to the multistage adaptive design

As noted earlier, the MSAT design for mathematics expanded and enhanced what was accomplished with the adaptive design for reading in PISA 2018. Test assembly for PISA 2022 was implemented in four steps:

1. Assemble non-overlapping parallel item sets.
2. Assemble core and adaptive testlets from each item set.
3. Assemble multistage adaptive paths using the core and adaptive testlets.
4. Assemble linear forms using the core testlets.

In each step, mixed-integer linear programming was used (van der Linden, 2005<sup>[3]</sup>; Diao and van der Linden, 2011<sup>[4]</sup>; van Rijn et al., 2022<sup>[5]</sup>). In the first step, the decision variables were defined as which unit was to be in which item set. For the second step, the decision variables were defined as which unit was to be in which testlet. In the third step, they were to describe which of the core and adaptive testlets was in which multistage adaptive path. Finally, in step four, they indicated which core testlets were in which linear form. Furthermore, all steps but the first consisted of multiple assemblies (e.g. in step 2, 16 core testlets were assembled from item set A, 16 core testlets were from item set B, etc.)

The objective in each step was always to minimize the difference with respect to a target TIF. In each step, constraints on the following variables were set: item exposure, number of units, number of items, maximum score, maximum human score, number of trend/new items, number of dichotomous/polytomous items, item format, content subdomain, process subdomain, overlap, median response time, and TCC.

As an example, the assembly of a set of core testlets is illustrated. In this case, the main decision variables of the assembly are defined as follows

$$x_{ut} = \begin{cases} 1, & \text{if unit } u \text{ in testlet } t, \\ 0, & \text{otherwise.} \end{cases}$$

Formula 2.1

Under local independence, the information function of a unit is the sum of item information functions :  $I_u(\theta) = \sum_{i \in V_u} I_i(\theta)$ , where  $V_u$  indicates the set of items in unit  $u$ . Similarly, the unit characteristic curve (i.e. the expected score on a unit as a function of  $\theta$ ) is the sum of item characteristic curves :  $T_u(\theta) = \sum_{i \in V_u} T_i(\theta)$ . The target TIF is denoted by  $J(\theta)$  and the objective is to minimize  $\epsilon$  subject to

$$J(\theta_j) - \epsilon \leq \sum_{u=1}^U I_u(\theta_j) x_{ut} \leq J(\theta_j) + \epsilon, \quad \text{for all } j \text{ and } t,$$

Formula 2.2

where  $\epsilon > 0$  and  $U$  is the number of units in the used item set. For the core testlets, the target TIF was set proportional to the TIF of the item set. The number of  $\theta$  points, indexed by  $j$ , at which to evaluate the TIF was three. To avoid potential differences in the TCC, an interval of one score point around the target TCC,  $J(\theta_j)$ , was allowed, which can be formalized as

$$J(\theta_j) - 0.5 \leq \sum_{u=1}^U T_u(\theta_j) x_{ut} \leq J(\theta_j) + 0.5, \quad \text{for all } j \text{ and } t.$$

Formula 2.3

Other constraints of category  $c$  can be formulated as:

$$n_c^{\min} \leq \sum_{u=1}^U n_{cu} x_{ut} \leq n_c^{\max}, \quad \text{for all } t,$$

Formula 2.4

where  $n_c^{\min}$  is the minimum required number (e.g. the number of items, the maximum score),  $n_{cu}$  is the number for category  $c$  of unit  $u$ , and  $n_c^{\max}$  is the maximum required number. Note that the constraints here can be both categorical and numerical. For the core testlets, the number of items was constrained to either 9 or 10 and the maximum score to 12 or 13. Bounds on the number of common items between testlets (overlap) can be added with the following set of constraints:

$$\begin{aligned} n_o^{\min} &\leq \sum_{u=1}^U n_u z_{utt'} \leq n_o^{\max}, & \text{for all } t < t', \\ 2z_{utt'} &\leq x_{ut} + x_{ut'}, & \text{for all } u, \\ z_{utt'} &\geq x_{ut} + x_{ut'} - 1, & \text{for all } u, \end{aligned}$$

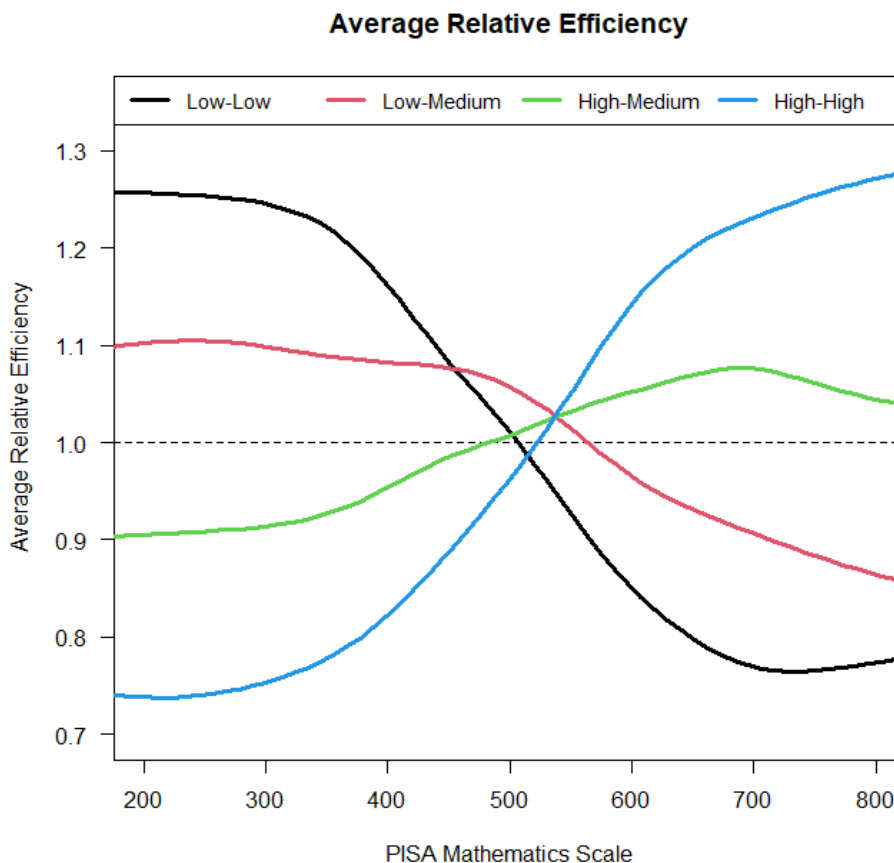
Formula 2.5

where  $n_o^{\min}$  and  $n_o^{\max}$  are the minimum and maximum number of common items,  $n_u$  is the number of items of unit  $u$ ,  $z_{utt'}$  are additional decision variables that indicate whether unit  $u$  is in both testlet  $t$  and  $t'$  with  $t < t'$ . The last two constraints are needed to keep the values of the decision variables consistent [see van der Linden (2005, p. 145<sub>[3]</sub>)].

Across all steps of the assembly, the total number of decision variables was about 92,000 and the total number of constraints was about 174,000, too many to list all of them here. Furthermore, the assembly was an iterative process in the sense that desired constraints could not always be implemented due to availability (e.g. not enough items of a specific type) or infeasibility (i.e. no solution could be found). In the latter case, a process called feasibility relaxation was used in which weights were assigned to give higher priority to more problematic constraint violations (e.g. items being overused) and lower priority to less problematic constraint violations (e.g. content constraints) [e.g. Lundell and Kronqvist (2022<sub>[6]</sub>)].

To evaluate the expected efficiency of the MSAT design, Figure 2.5 shows the average relative efficiency based on the average TIF of the MSAT paths over the average TIF of linear forms using estimated item parameters from the field trial (only international item parameters were used). Values larger than one indicate that the MSAT paths are more efficient than the linear forms. It can be seen that the MSAT paths provide more information than the linear forms when the proficiency level would match the difficulty level of the path (e.g. the curve for the low-low path exceeds one for lower proficiency values).

**Figure 2.5. Average relative efficiency of MSAT paths over linear forms for PISA 2022 mathematics test design**



## References

- Ali, U. and P. van Rijn (2016), “An evaluation of different statistical targets for assembling parallel forms in item response theory”, *Applied Psychological Measurement*, Vol. 40/3, pp. 163-179, <https://doi.org/10.1177/0146621615613308>. [2]
- Diao, Q. and W. van der Linden (2011), “Automated test assembly using lp\_solve version 5.5 in R”, *Applied Psychological Measurement*, Vol. 35/5, pp. 398-409. [4]
- Lundell, A. and J. Kronqvist (2022), “Polyhedral approximation strategies for nonconvex mixed-integer nonlinear programming in SHOT”, *Journal of Global Optimization*, Vol. 82/4, pp. 863-896. [6]
- van der Linden, W. (2005), *Linear Models for Optimal Test Design*, Springer, New York. [3]
- van Rijn, P. et al. (2022), *Stepwise assembly for multistage adaptive testing: An application to PISA mathematics*, Presentation at IACAT conference, Frankfurt, Germany. [5]
- von Davier, M. et al. (2019), “Evaluating item response theory linking and model fit for data from PISA 2000-2012”, *Assessment in Education: Principles, Policy & Practice*, Vol. 26/4, pp. 466-488. [1]

## Notes

- 
1. The mode of assessment for most of the participants was computer-based (77 CBA participants), with 4 participants implementing the PISA 2022 cycle as a paper-based survey.
  2. See <https://www.oecd.org/pisa/pisa-for-development/>.

## Annex 2.A. Main survey items

Annex Table 2.A.1. Chapter 2: Main survey mathematics analysis

Tables	Title
Table 2.A.2	Number of items by domain and across cycles in the main survey
Table 2.A.3	Main survey computer-based MSAT paths for mathematics
Table 2.A.4	Main survey computer-based linear forms for mathematics

Annex Table 2.A.2. Number of items by domain and across cycles in the main survey

	2000	2003	2006	2009	2012	2015	2018	2022
Reading	<b>129</b>	28	28	<b>131</b>	44	103	<b>245</b>	197
Mathematics	43	<b>84</b>	48	35	<b>109</b>	83	83	<b>234</b>
Science	45	34	<b>103</b>	53	53	<b>184</b>	115	115
Total Across Domains	217	146	179	219	206	370	443	546

Note: Bold numbers indicate the major domain in each cycle. For the 2015 and 2018 cycles, the computer-based mathematics instrument contained 82 items, while the equivalent paper-based instrument contained 83 items. This is because there was one item that was not able to be transitioned to a computer-based delivery in 2015 (the item requires students to draw on a map). The number of mathematics items in the 2022 cycle includes 74 "trend" items (i.e. items developed prior to this cycle) and 160 "new" items (i.e. items developed this cycle).

Annex Table 2.A.3. Main survey computer-based MSAT paths for mathematics

MSAT_Path	Difficulty_Level	Core_Testlet	Adaptive_1_Testlet	Adaptive_2_Testlet
1	Low_Low	MTA001	MTB108	MTC203
2	Low_Low	MTA002	MTB103	MTC204
3	Low_Low	MTA003	MTB105	MTC204
4	Low_Low	MTA004	MTB105	MTC201
5	Low_Low	MTA005	MTB104	MTC202
6	Low_Low	MTA006	MTB107	MTC204
7	Low_Low	MTA007	MTB104	MTC202
8	Low_Low	MTA008	MTB108	MTC203
9	Low_Low	MTA009	MTB101	MTC201
10	Low_Low	MTA010	MTB106	MTC202
11	Low_Low	MTA011	MTB101	MTC201
12	Low_Low	MTA012	MTB103	MTC203
13	Low_Low	MTA013	MTB107	MTC202
14	Low_Low	MTA014	MTB102	MTC201
15	Low_Low	MTA015	MTB102	MTC203
16	Low_Low	MTA016	MTB106	MTC204
17	Low_Medium	MTA001	MTB108	MTC206
18	Low_Medium	MTA002	MTB103	MTC212
19	Low_Medium	MTA003	MTB105	MTC205
20	Low_Medium	MTA004	MTB105	MTC208
21	Low_Medium	MTA005	MTB104	MTC211
22	Low_Medium	MTA006	MTB107	MTC206
23	Low_Medium	MTA007	MTB104	MTC207

MSAT_Path	Difficulty_Level	Core_Testlet	Adaptive_1_Testlet	Adaptive_2_Testlet
24	Low_Medium	MTA008	MTB108	MTC210
25	Low_Medium	MTA009	MTB101	MTC208
26	Low_Medium	MTA010	MTB106	MTC210
27	Low_Medium	MTA011	MTB101	MTC212
28	Low_Medium	MTA012	MTB103	MTC209
29	Low_Medium	MTA013	MTB107	MTC211
30	Low_Medium	MTA014	MTB102	MTC209
31	Low_Medium	MTA015	MTB102	MTC205
32	Low_Medium	MTA016	MTB106	MTC207
33	High_Medium	MTA001	MTB113	MTC206
34	High_Medium	MTA002	MTB115	MTC212
35	High_Medium	MTA003	MTB110	MTC205
36	High_Medium	MTA004	MTB112	MTC211
37	High_Medium	MTA005	MTB116	MTC207
38	High_Medium	MTA006	MTB111	MTC209
39	High_Medium	MTA007	MTB114	MTC211
40	High_Medium	MTA008	MTB114	MTC210
41	High_Medium	MTA009	MTB113	MTC208
42	High_Medium	MTA010	MTB110	MTC210
43	High_Medium	MTA011	MTB115	MTC212
44	High_Medium	MTA012	MTB109	MTC206
45	High_Medium	MTA013	MTB116	MTC207
46	High_Medium	MTA014	MTB109	MTC205
47	High_Medium	MTA015	MTB111	MTC209
48	High_Medium	MTA016	MTB112	MTC208
49	High_High	MTA001	MTB113	MTC213
50	High_High	MTA002	MTB115	MTC214
51	High_High	MTA003	MTB110	MTC216
52	High_High	MTA004	MTB112	MTC213
53	High_High	MTA005	MTB116	MTC215
54	High_High	MTA006	MTB111	MTC214
55	High_High	MTA007	MTB114	MTC215
56	High_High	MTA008	MTB114	MTC214
57	High_High	MTA009	MTB113	MTC213
58	High_High	MTA010	MTB110	MTC216
59	High_High	MTA011	MTB115	MTC214
60	High_High	MTA012	MTB109	MTC215
61	High_High	MTA013	MTB116	MTC215
62	High_High	MTA014	MTB109	MTC216
63	High_High	MTA015	MTB111	MTC213
64	High_High	MTA016	MTB112	MTC216
65	Low_Low	MTB001	MTC103	MTA204
66	Low_Low	MTB002	MTC107	MTA201
67	Low_Low	MTB003	MTC101	MTA204
68	Low_Low	MTB004	MTC106	MTA203
69	Low_Low	MTB005	MTC104	MTA201
70	Low_Low	MTB006	MTC103	MTA204
71	Low_Low	MTB007	MTC105	MTA203
72	Low_Low	MTB008	MTC102	MTA203
73	Low_Low	MTB009	MTC108	MTA202
74	Low_Low	MTB010	MTC106	MTA201
75	Low_Low	MTB011	MTC108	MTA202
76	Low_Low	MTB012	MTC107	MTA201

MSAT_Path	Difficulty_Level	Core_Testlet	Adaptive_1_Testlet	Adaptive_2_Testlet
77	Low_Low	MTB013	MTC105	MTA203
78	Low_Low	MTB014	MTC104	MTA202
79	Low_Low	MTB015	MTC101	MTA204
80	Low_Low	MTB016	MTC102	MTA202
81	Low_Medium	MTB001	MTC103	MTA212
82	Low_Medium	MTB002	MTC107	MTA206
83	Low_Medium	MTB003	MTC101	MTA211
84	Low_Medium	MTB004	MTC106	MTA208
85	Low_Medium	MTB005	MTC104	MTA205
86	Low_Medium	MTB006	MTC103	MTA211
87	Low_Medium	MTB007	MTC105	MTA208
88	Low_Medium	MTB008	MTC102	MTA209
89	Low_Medium	MTB009	MTC108	MTA206
90	Low_Medium	MTB010	MTC106	MTA207
91	Low_Medium	MTB011	MTC108	MTA209
92	Low_Medium	MTB012	MTC107	MTA207
93	Low_Medium	MTB013	MTC105	MTA205
94	Low_Medium	MTB014	MTC104	MTA210
95	Low_Medium	MTB015	MTC101	MTA212
96	Low_Medium	MTB016	MTC102	MTA210
97	High_Medium	MTB001	MTC113	MTA211
98	High_Medium	MTB002	MTC114	MTA208
99	High_Medium	MTB003	MTC112	MTA212
100	High_Medium	MTB004	MTC113	MTA212
101	High_Medium	MTB005	MTC110	MTA205
102	High_Medium	MTB006	MTC114	MTA208
103	High_Medium	MTB007	MTC109	MTA211
104	High_Medium	MTB008	MTC115	MTA207
105	High_Medium	MTB009	MTC110	MTA209
106	High_Medium	MTB010	MTC111	MTA209
107	High_Medium	MTB011	MTC115	MTA206
108	High_Medium	MTB012	MTC116	MTA210
109	High_Medium	MTB013	MTC109	MTA206
110	High_Medium	MTB014	MTC116	MTA210
111	High_Medium	MTB015	MTC112	MTA205
112	High_Medium	MTB016	MTC111	MTA207
113	High_High	MTB001	MTC113	MTA213
114	High_High	MTB002	MTC114	MTA215
115	High_High	MTB003	MTC112	MTA216
116	High_High	MTB004	MTC113	MTA214
117	High_High	MTB005	MTC110	MTA214
118	High_High	MTB006	MTC114	MTA215
119	High_High	MTB007	MTC109	MTA216
120	High_High	MTB008	MTC115	MTA214
121	High_High	MTB009	MTC110	MTA213
122	High_High	MTB010	MTC111	MTA215
123	High_High	MTB011	MTC115	MTA216
124	High_High	MTB012	MTC116	MTA213
125	High_High	MTB013	MTC109	MTA216
126	High_High	MTB014	MTC116	MTA215
127	High_High	MTB015	MTC112	MTA214
128	High_High	MTB016	MTC111	MTA213
129	Low_Low	MTC001	MTA103	MTB201



MSAT_Path	Difficulty_Level	Core_Testlet	Adaptive_1_Testlet	Adaptive_2_Testlet
130	Low_Low	MTC002	MTA101	MTB202
131	Low_Low	MTC003	MTA107	MTB202
132	Low_Low	MTC004	MTA105	MTB201
133	Low_Low	MTC005	MTA101	MTB204
134	Low_Low	MTC006	MTA104	MTB201
135	Low_Low	MTC007	MTA108	MTB202
136	Low_Low	MTC008	MTA104	MTB203
137	Low_Low	MTC009	MTA103	MTB203
138	Low_Low	MTC010	MTA102	MTB203
139	Low_Low	MTC011	MTA106	MTB204
140	Low_Low	MTC012	MTA107	MTB202
141	Low_Low	MTC013	MTA106	MTB204
142	Low_Low	MTC014	MTA105	MTB201
143	Low_Low	MTC015	MTA108	MTB204
144	Low_Low	MTC016	MTA102	MTB203
145	Low_Medium	MTC001	MTA103	MTB210
146	Low_Medium	MTC002	MTA101	MTB212
147	Low_Medium	MTC003	MTA107	MTB207
148	Low_Medium	MTC004	MTA105	MTB211
149	Low_Medium	MTC005	MTA101	MTB208
150	Low_Medium	MTC006	MTA104	MTB209
151	Low_Medium	MTC007	MTA108	MTB211
152	Low_Medium	MTC008	MTA104	MTB208
153	Low_Medium	MTC009	MTA103	MTB210
154	Low_Medium	MTC010	MTA102	MTB205
155	Low_Medium	MTC011	MTA106	MTB206
156	Low_Medium	MTC012	MTA107	MTB209
157	Low_Medium	MTC013	MTA106	MTB212
158	Low_Medium	MTC014	MTA105	MTB205
159	Low_Medium	MTC015	MTA108	MTB207
160	Low_Medium	MTC016	MTA102	MTB206
161	High_Medium	MTC001	MTA112	MTB205
162	High_Medium	MTC002	MTA115	MTB210
163	High_Medium	MTC003	MTA110	MTB209
164	High_Medium	MTC004	MTA113	MTB207
165	High_Medium	MTC005	MTA109	MTB206
166	High_Medium	MTC006	MTA109	MTB212
167	High_Medium	MTC007	MTA112	MTB208
168	High_Medium	MTC008	MTA114	MTB207
169	High_Medium	MTC009	MTA115	MTB208
170	High_Medium	MTC010	MTA110	MTB205
171	High_Medium	MTC011	MTA113	MTB211
172	High_Medium	MTC012	MTA116	MTB209
173	High_Medium	MTC013	MTA114	MTB206
174	High_Medium	MTC014	MTA116	MTB210
175	High_Medium	MTC015	MTA111	MTB212
176	High_Medium	MTC016	MTA111	MTB211
177	High_High	MTC001	MTA112	MTB215
178	High_High	MTC002	MTA115	MTB214
179	High_High	MTC003	MTA110	MTB216
180	High_High	MTC004	MTA113	MTB214
181	High_High	MTC005	MTA109	MTB216
182	High_High	MTC006	MTA109	MTB215

MSAT_Path	Difficulty_Level	Core_Testlet	Adaptive_1_Testlet	Adaptive_2_Testlet
183	High_High	MTC007	MTA112	MTB213
184	High_High	MTC008	MTA114	MTB214
185	High_High	MTC009	MTA115	MTB214
186	High_High	MTC010	MTA110	MTB216
187	High_High	MTC011	MTA113	MTB213
188	High_High	MTC012	MTA116	MTB216
189	High_High	MTC013	MTA114	MTB213
190	High_High	MTC014	MTA116	MTB215
191	High_High	MTC015	MTA111	MTB213
192	High_High	MTC016	MTA111	MTB215

Where:

MT = Math Testlet

A-B-C = Set

0-1-2 (the single digit immediately to the right of the set letter) = Stage

0 = core, 1 = adaptive stage 1, 2 = adaptive stage 2

01-16 (the last two digits on the right) = testlet number

Examples:

MTA008	<b>MT</b>	<b>A</b>	<b>0</b>	<b>08</b>
	Math Testlet	set A	core stage	testlet 08
MTB214	<b>MT</b>	<b>B</b>	<b>2</b>	<b>14</b>
	Math Testlet	set B	adaptive stage 2	testlet 14

#### Annex Table 2.A.4. Main survey computer-based linear forms for mathematics

Linear_Form	Core_Testlet_1	Core_Testlet_2	Core_Testlet_3
1	MTA001	MTB010	MTC015
2	MTA002	MTB014	MTC010
3	MTA003	MTB001	MTC013
4	MTA004	MTB011	MTC014
5	MTA005	MTB005	MTC005
6	MTA006	MTB008	MTC004
7	MTA007	MTB016	MTC012
8	MTA008	MTB007	MTC016
9	MTA009	MTB006	MTC009
10	MTA010	MTB009	MTC011
11	MTA011	MTB004	MTC008
12	MTA012	MTB013	MTC006
13	MTA013	MTB002	MTC002
14	MTA014	MTB003	MTC001
15	MTA015	MTB012	MTC007
16	MTA016	MTB015	MTC003
17	MTB001	MTC001	MTA010
18	MTB002	MTC005	MTA002
19	MTB003	MTC008	MTA004
20	MTB004	MTC003	MTA015
21	MTB005	MTC014	MTA007
22	MTB006	MTC010	MTA011
23	MTB007	MTC016	MTA005
24	MTB008	MTC006	MTA012
25	MTB009	MTC011	MTA014

Linear_Form	Core_Testlet_1	Core_Testlet_2	Core_Testlet_3
26	MTB010	MTC015	MTA001
27	MTB011	MTC013	MTA016
28	MTB012	MTC002	MTA009
29	MTB013	MTC009	MTA006
30	MTB014	MTC007	MTA003
31	MTB015	MTC004	MTA008
32	MTB016	MTC012	MTA013
33	MTC001	MTA012	MTB002
34	MTC002	MTA005	MTB007
35	MTC003	MTA015	MTB013
36	MTC004	MTA008	MTB015
37	MTC005	MTA002	MTB004
38	MTC006	MTA006	MTB008
39	MTC007	MTA016	MTB012
40	MTC008	MTA004	MTB011
41	MTC009	MTA010	MTB006
42	MTC010	MTA009	MTB009
43	MTC011	MTA014	MTB014
44	MTC012	MTA013	MTB016
45	MTC013	MTA007	MTB003
46	MTC014	MTA003	MTB005
47	MTC015	MTA001	MTB010
48	MTC016	MTA011	MTB001

Where:

MT = Math Testlet

A-B-C = Set

0 (the single digit immediately to the right of the set letter) = Stage

Only core testlets were used with the linear design

01-16 (the last two digits on the right) = testlet number

## Annex 2.B. Core testlet

Annex Table 2.B.1. Core testlet cut scores

Core Testlet	Core Cut Score Low-High	Max. Machine Score	Max. Total Score
MTA001	6	10	13
MTA002	5	11	13
MTA003	6	11	13
MTA004	6	9	12
MTA005	6	11	13
MTA006	6	9	13
MTA007	6	10	12
MTA008	6	9	12
MTA009	5	11	13
MTA010	5	11	13
MTA011	5	10	13
MTA012	6	11	13
MTA013	5	9	12
MTA014	6	11	13
MTA015	6	11	13
MTA016	6	9	12
MTB001	6	9	12
MTB002	6	9	13
MTB003	6	10	12
MTB004	6	11	13
MTB005	6	11	13
MTB006	6	9	12
MTB007	6	9	12
MTB008	5	11	13
MTB009	6	10	13
MTB010	6	11	13
MTB011	6	11	13
MTB012	6	11	13
MTB013	5	11	13
MTB014	6	11	13
MTB015	6	9	12
MTB016	6	10	12
MTC001	6	10	12
MTC002	6	11	13
MTC003	6	11	13
MTC004	5	11	13

Core Testlet	Core Cut Score Low-High	Max. Machine Score	Max. Total Score
MTC005	6	10	12
MTC006	6	11	13
MTC007	6	9	11
MTC008	6	12	14
MTC009	5	11	13
MTC010	6	10	12
MTC011	5	10	12
MTC012	5	9	13
MTC013	5	10	12
MTC014	6	10	13
MTC015	5	10	12
MTC016	6	11	13

## Annex 2.C. Adaptive testlet

Annex Table 2.C.1. Adaptive testlet cut scores

Core Testlet	Adaptive 1 Testlet	Adaptive 1 Cut Score Low-Medium	Adaptive 1 Cut Score Medium-High	Max. Machine Score	Max. Total Score
MTA001	MTB108	9	99	20	26
MTA001	MTB113	99	12	19	25
MTA002	MTB103	9	99	21	25
MTA002	MTB115	99	12	20	25
MTA003	MTB105	10	99	21	25
MTA003	MTB110	99	12	21	26
MTA004	MTB105	10	99	19	24
MTA004	MTB112	99	11	19	24
MTA005	MTB104	11	99	22	26
MTA005	MTB116	99	12	20	25
MTA006	MTB107	10	99	20	26
MTA006	MTB111	99	13	20	26
MTA007	MTB104	11	99	21	25
MTA007	MTB114	99	12	21	25
MTA008	MTB108	9	99	19	25
MTA008	MTB114	99	12	20	25
MTA009	MTB101	10	99	22	26
MTA009	MTB113	99	11	20	25
MTA010	MTB106	10	99	21	26
MTA010	MTB110	99	11	21	26
MTA011	MTB101	10	99	21	26
MTA011	MTB115	99	12	19	25
MTA012	MTB103	10	99	21	25
MTA012	MTB109	99	14	22	26
MTA013	MTB107	10	99	20	25
MTA013	MTB116	99	11	18	24
MTA014	MTB102	10	99	21	25
MTA014	MTB109	99	13	22	26
MTA015	MTB102	11	99	21	25
MTA015	MTB111	99	13	22	26
MTA016	MTB106	11	99	19	25
MTA016	MTB112	99	11	19	24
MTB001	MTC103	10	99	19	24
MTB001	MTC113	99	12	20	25
MTB002	MTC107	11	99	20	26
MTB002	MTC114	99	12	19	25
MTB003	MTC101	10	99	20	24

Core Testlet	Adaptive 1 Testlet	Adaptive 1 Cut Score Low-Medium	Adaptive 1 Cut Score Medium-High	Max. Machine Score	Max. Total Score
MTB003	MTC112	99	12	20	24
MTB004	MTC106	9	99	21	26
MTB004	MTC113	99	12	22	26
MTB005	MTC104	11	99	22	26
MTB005	MTC110	99	13	21	25
MTB006	MTC103	10	99	19	24
MTB006	MTC114	99	12	19	24
MTB007	MTC105	10	99	19	25
MTB007	MTC109	99	12	19	25
MTB008	MTC102	9	99	22	26
MTB008	MTC115	99	11	21	25
MTB009	MTC108	10	99	20	25
MTB009	MTC110	99	13	20	25
MTB010	MTC106	9	99	21	26
MTB010	MTC111	99	11	21	26
MTB011	MTC108	10	99	21	25
MTB011	MTC115	99	13	21	25
MTB012	MTC107	11	99	22	26
MTB012	MTC116	99	12	21	26
MTB013	MTC105	10	99	21	26
MTB013	MTC109	99	12	21	26
MTB014	MTC104	11	99	22	26
MTB014	MTC116	99	13	21	26
MTB015	MTC101	10	99	19	24
MTB015	MTC112	99	12	19	24
MTB016	MTC102	10	99	21	25
MTB016	MTC111	99	11	20	25
MTC001	MTA103	10	99	20	24
MTC001	MTA112	99	13	21	25
MTC002	MTA101	10	99	21	25
MTC002	MTA115	99	13	21	25
MTC003	MTA107	9	99	21	27
MTC003	MTA110	99	12	21	26
MTC004	MTA105	9	99	22	26
MTC004	MTA113	99	12	22	26
MTC005	MTA101	9	99	20	24
MTC005	MTA109	99	12	21	25
MTC006	MTA104	11	99	21	25
MTC006	MTA109	99	13	22	26
MTC007	MTA108	10	99	19	23
MTC007	MTA112	99	13	20	24
MTC008	MTA104	10	99	22	26
MTC008	MTA114	99	12	22	27

Core Testlet	Adaptive 1 Testlet	Adaptive 1 Cut Score Low-Medium	Adaptive 1 Cut Score Medium-High	Max. Machine Score	Max. Total Score
MTC009	MTA103	9	99	21	25
MTC009	MTA115	99	12	21	25
MTC010	MTA102	9	99	19	25
MTC010	MTA110	99	12	20	25
MTC011	MTA106	10	99	21	25
MTC011	MTA113	99	12	21	25
MTC012	MTA107	9	99	19	27
MTC012	MTA116	99	11	21	25
MTC013	MTA106	10	99	21	25
MTC013	MTA114	99	12	20	25
MTC014	MTA105	11	99	21	26
MTC014	MTA116	99	13	22	25
MTC015	MTA108	9	99	20	24
MTC015	MTA111	99	12	21	25
MTC016	MTA102	9	99	20	26
MTC016	MTA111	99	13	22	26

Note: 99 = not applicable.



# 3 Test Development for the Core Domains

## Introduction

This chapter describes the processes used by the PISA Core A contractor, Educational Testing Service (ETS), and the international test development team to develop the tests for the core domains in the PISA 2022 cycle.

The tests for the PISA 2022 cycle included the following:

- a mathematics test, the major domain in PISA 2022
- a reading and a science test, the two minor domains
- a creative thinking test, the innovative domain for this cycle, and
- a financial literacy test, the international option for this cycle.

Test design and development for the Creative Thinking domain is presented and discussed in the Chapter 4 [*Development of the PISA 2022 Innovative Domain Assessment*] of this technical report.

In the PISA 2015 cycle, PISA moved 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 participants that chose not to implement the computer-delivered survey. The PISA 2018 cycle retained this same paper-based option, using the same paper-based materials as the PISA 2015 cycle. The PISA 2022 cycle retained this paper-based option as well; however, only one participant used the same paper-based materials as in the 2015 and 2018 cycles. The other paper-based participants administered a “new” instrument that was first used in the PISA for Development (PISA-D) assessment. This “new” paper-based instrument, which contained a substantial amount of material that overlapped with computer- and paper-based trend material administered by other participants, was comprised of clusters of units assessing mathematics, science, reading, and reading components.

The computer-based delivery mode allows PISA to measure new and expanded aspects of the domain constructs. In mathematics, new material for PISA 2022 included items developed to assess mathematical reasoning as a separate process classification, and items that leveraged the use of the digital environment (e.g. spreadsheets, simulators, data generators, drag-and-drop, etc.). A mixed-design that included a computer-based multistage adaptive testing was also adopted for the mathematics literacy domain to further improve measurement accuracy and efficiency, especially at the extremes of the proficiency scale. In financial literacy, some new units were developed based on an updated framework and to ensure adequate coverage of the domain following the release/removal of several units following the 2018 administration.

As noted in the list above, the core domains in PISA rotate between being a major or a minor domain. Annex Table 3.A.2. shows the number of items in the main survey for the core domains for each PISA cycle since PISA 2000. Under this approach for measuring trends, each domain goes through a domain

rotation that begins with a new or revised framework and continues with the two subsequent cycles in which it becomes a minor domain. The rotation concludes, and starts again, with becoming a major domain three cycles later. The third cycle- after alternating with the other two main domains - then involves another revision of the framework to reflect the current thinking about assessment for the new data collection as a major domain. For example, the revised framework for mathematics as the major domain in PISA 2022 and the introduction of computer-based items broadened the construct beyond what was measured in PISA 2012, the last time that mathematics was a major domain. Under the current design, the mathematics framework and instruments are expected to remain constant for the next two PISA cycles, with the next revision of the mathematics assessment and items expected for the PISA cycle to take place after PISA 2029, when mathematics will again be the major domain. Note that over time, the number of items included for minor domains has increased, which has helped stabilize and improve the measurement of trends for the minor domains by making the construct coverage for each minor domain comparable to that of a major domain. However, there has been a reduction in the number of student responses per item for the minor domains.

In addition to the three core domains (science, mathematics, and reading) and the innovative domain (creative thinking), the PISA 2022 assessment also included an optional assessment of financial literacy, which was administered only as a computer-based assessment.

Annex Table 3.A.3 and Annex Table 3.A.4 present the domain coverage for the computer- and paper-based assessments, respectively. All new items for mathematics were developed as computer-based items. The mathematics field trial design included seven clusters of trend items and twelve clusters of new items to study unit order effects. This was carried out in preparation for the introduction of the multistage adaptive testing design in the main survey. Then, in the main survey, the mathematics items were assigned according to the multistage adaptive design described in Chapter 2 [*The PISA 2022 Integrated Assessment Design*] of this report.

As shown in Annex Table 3.A.3, there was no new item development for science or reading in PISA 2022. Both financial literacy and creative thinking were administered only as part of the computer-based assessment and therefore all item development was done for computer delivery, although most of the trend items for financial literacy were originally developed for a paper-based administration.

As shown in Annex Table 3.A.4, there was a paper-based instrument that was used in the PISA 2015 and the PISA 2018 cycles, which contain only items taken from cycles prior to PISA 2015. Only one of the participants administered these instruments. The other three paper-based participants used a “new” paper-based instrument that was first used in PISA for Development.

### ***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 trend items (i.e. items developed prior to PISA 2015) in each core domain and had a reduced reading load. The UH form contained about half as many items as the other forms, with each cluster including from seven to nine items. In PISA 2022 the UH form was comprised of about 53% mathematics, 21% reading, and 26% science items. The UH form included two 15-minute clusters of mathematics (MU1 and MU2), one 15-minute cluster of reading (RU1) and one 15-minute cluster of science (SU1).

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

## Assessment of financial literacy

The assessment of financial literacy was again offered as an international option in PISA 2022. The financial literacy instrument included trend items from the PISA 2012, PISA 2015, and PISA 2018 assessments, plus a few new units that were developed for PISA 2022. Financial literacy was administered only as a computer-based assessment.

Like in PISA 2018, the financial literacy assessment was administered to a separate sample of PISA-eligible students who took, in addition to the financial literacy assessment, reading or mathematics items. As with students sitting PISA as part of the main sample described in Chapter 2, the total testing time for each student was two hours (120 minutes) for the cognitive test.

## The 2022 mathematics assessment framework<sup>1</sup>

For each PISA domain, an assessment framework is created 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 target distribution of items according to important framework dimensions – and outline the possibilities for reporting results. For PISA 2022, a subject matter expert group (SMEG) was convened to develop a framework for mathematical literacy under the guidance of RTI International and with input from the PISA Governing Board and Core A (ETS). A separate expert group, convened by ACT (Core B3), worked on creative thinking.

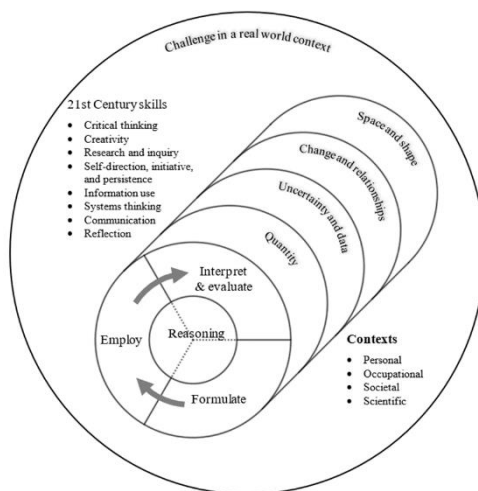
Mathematical literacy, for PISA 2022, is defined as follows:

*Mathematical literacy is an individual's capacity to reason mathematically and to formulate, employ, and interpret mathematics to solve problems in a variety of real-world contexts. It includes concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to know the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective 21st century citizens.*

Additionally, the definition of mathematical literacy for PISA 2022 can be considered with respect to three interrelated concepts, which are represented in Figure 3.1 and will be explained in this section. These interrelated concepts are:

1. **Cognitive Processes:** mathematical reasoning and the problem-solving model
2. **Content Knowledge:** how the domain is organized into categories
3. **Contexts:** the real-world “setting” in which items are presented, including select 21<sup>st</sup> Century skills that are supported and developed as part of being mathematically literate.

Figure 3.1. Mathematical literacy for PISA 2022



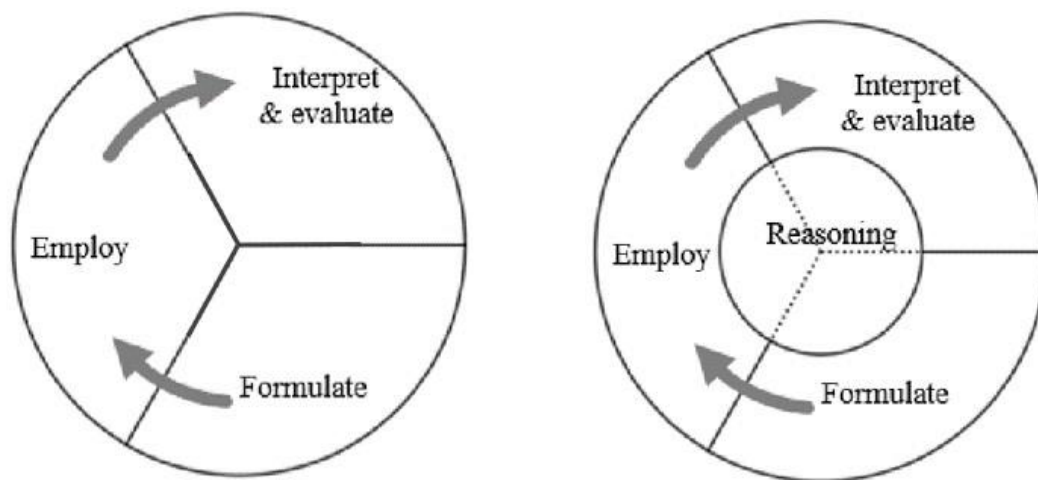
### ***Mathematical Literacy – Cognitive Processes***

For PISA 2022, the mathematical literacy domain describes mathematics in terms of four cognitive processes: reasoning, formulating, employing, and interpreting/evaluating.

Previous PISA mathematics frameworks used three cognitive processes (formulating, employing, and interpreting/evaluating), which formed the basis of the mathematical problem-solving model. For PISA 2022, reasoning was included as a separate cognitive process, but it is not a new concept in PISA mathematics. Reasoning – including both deductive (i.e. mathematical) and inductive (i.e. statistical) reasoning – has always existed as an underlying element to the problem-solving model and is considered a core aspect of being mathematically literate; therefore, the updated mathematics framework sought to highlight reasoning as both a central component underlying the processes in the problem-solving model, and as its own process.

Figure 3.2 shows the mathematical problem-solving model used in previous cycles and in the current cycle with reasoning as a fourth process. Note that even though the problem-solving model is comprised of multiple processes, each PISA mathematics item was written specifically towards one of the processes and students were not expected to utilize the full model to respond to each item. For example, a formulate item might assess if a student can write an equation to model a situation without requiring application of any processes/procedures (i.e. employing) or reflection on the result (i.e. interpreting/evaluating). The cognitive processes within each category are briefly defined below.

Figure 3.2. Cognitive processes and the mathematical problem-solving model: prior to 2022 (left) and for 2022 (right)



### *Reasoning Mathematically*

- *Reasoning* mathematically (both deductively and inductively) involves evaluating situations, selecting strategies, drawing logical conclusions, developing and describing solutions, and recognising how those solutions can be applied. Students reason mathematically when they identify, recognise, organise, connect, and represent

:

- Construct, abstract, evaluate, deduce, justify, explain, and defend
- Interpret, make judgements, critique, refute, and qualify

### *Formulating Situations Mathematically*

*Formulating* situations mathematically refers to individuals being able to recognise and identify opportunities to use mathematics and then providing mathematical structure to a problem presented in some contextualised form, including reasoning about the constraints and assumptions in the problem, which may involve:

- Selecting an appropriate model from a list
- Identifying the mathematical aspects of a problem situated in a real-world context and identifying the significant variables
- Recognising mathematical structure (including regularities, relationships, and patterns) in problems or situations
- Simplifying a situation or problem in order to make it amenable to mathematical analysis (for example by decomposing)
- Identifying constraints and assumptions behind any mathematical modelling and simplifications gleaned from the context
- Representing a situation mathematically, using appropriate variables, symbols, diagrams, and standard models
- Representing a problem in a different way, including organising it according to mathematical concepts and making appropriate assumptions

- Understanding and explaining the relationships between the context-specific language of a problem and the symbolic and formal language needed to represent it mathematically
- Translating a problem into mathematical language or a representation
- Recognising aspects of a problem that correspond with known problems or mathematical concepts, facts or procedures
- Choosing among an array of and employing the most effective computing tool to portray a mathematical relationship inherent in a contextualised problem
- Creating an ordered series of (step-by-step) instructions for solving problems.

### *Employing Mathematical Concepts, Facts, and Procedures*

*Employing* mathematical concepts, facts, and procedures refers to individuals being able to apply mathematical concepts, facts, procedures, and reasoning to solve mathematically formulated problems to obtain mathematical conclusions, including:

- Performing a simple calculation
- Drawing a simple conclusion
- Selecting an appropriate strategy from a list
- Devising and implementing strategies for finding mathematical solutions
- Using mathematical tools, including technology, to help find exact or approximate solutions
- Applying mathematical facts, rules, algorithms, and structures when finding solutions
- Manipulating numbers, graphical and statistical data and information, algebraic expressions and equations, and geometric representations
- Making mathematical diagrams, graphs, simulations, and constructions and extracting mathematical information from them
- Using and switching between different representations in the process of finding solutions
- Making generalisations and conjectures based on the results of applying mathematical procedures to find solutions
- Reflecting on mathematical arguments and explaining and justifying mathematical results
- Evaluating the significance of observed (or proposed) patterns and regularities in data

### *Interpreting, Applying and Evaluating Mathematical Outcomes*

- Interpreting, applying and evaluating mathematical outcomes refers to individuals being able to reflect upon mathematical solutions, results or conclusions and interpret them in the context of the real-life problem that initiated the process, including:
  - Interpreting information presented in graphical form and/or diagrams
  - Evaluating a mathematical outcome in terms of the context
  - Interpreting a mathematical result back into the real-world context
  - Evaluating the reasonableness of a mathematical solution in the context of a real-world problem
  - Understanding how the real world impacts the outcomes and calculations of a mathematical procedure or model in order to make contextual judgments about how the results should be adjusted or applied
- Explaining why a mathematical result or conclusion does, or does not, make sense given the context of a problem
- Understanding the extent and limits of mathematical concepts and mathematical solutions

- Critiquing and identifying the limits of the model used to solve a problem
- Using mathematical thinking and computational thinking to make predictions, to provide evidence for arguments, to test and compare proposed solutions.

### ***Mathematical Literacy – Content Knowledge***

The content of the PISA mathematics assessment is divided into the same four categories that were used in previous PISA cycles: quantity, uncertainty and data, change and relationships, and space and shape. Even though PISA is not a curriculum-driven assessment, these four categories reflect content that is common to many school curricula (i.e. content that most 15-year-olds are likely to have encountered in school) and cover a range of topics that are considered central to the study of mathematics.

A brief description of each of the four content categories is given below.

- **Quantity:** number sense and estimation; quantification of attributes, objects, relationships, situations and entities in the world; understanding various representations of those quantifications, and judging interpretations and arguments based on quantity.
- **Uncertainty and data:** recognising the place of variation in the real world, including having a sense of the quantification of that variation, and acknowledging its uncertainty and error in related inferences. It also includes forming, interpreting and evaluating conclusions drawn in situations where uncertainty is present. The presentation and interpretation of data are also included in this category, as well as basic topics in probability.
- **Change and relationships:** understanding fundamental types of change and recognising when they occur in order to use suitable mathematical models to describe and predict change. Includes appropriate functions and equations/inequalities as well as creating, interpreting and translating among symbolic and graphical representations of relationships.
- **Space and shape:** patterns; properties of objects; spatial visualisations; positions and orientations; representations of objects; decoding and encoding of visual information; navigation and dynamic interaction with real shapes as well as representations, movement, displacement, and the ability to anticipate actions in space.

Below is a list of content topics based on the results of an analysis of desired learning outcomes from a sample of eleven countries from around the world. These topics can be applied to one or more of the four content categories, and this list is not intended to be exhaustive, but rather reflective of content that is deemed important for students preparing to either enter the workforce or pursue higher levels of education. Additionally, mathematics experts have added a few focus topics pertinent to the updated framework.

- **Growth Phenomena:** different types of linear and non-linear growth
- **Geometric Approximation:** approximating the attributes and properties of irregular or unfamiliar shapes and objects by breaking these shapes and objects up into more familiar shapes and objects for which there are formulae and tools
- **Computer Simulations:** exploring situations (that may include budgeting, planning, population distribution, disease spread, experimental probability, reaction time modelling etc.) in terms of the variables and the impact that these have on the outcome
- **Conditional Decision-Making:** using basic principles of combinatorics and an understanding of interrelationships between variables to interpret situations and make predictions
- **Functions:** the concept of function, emphasising but not limited to linear functions, their properties, and a variety of descriptions and representations of them. Commonly used representations are verbal, symbolic, tabular and graphical.
- **Algebraic Expressions:** verbal interpretation of and manipulation with algebraic expressions, involving numbers, symbols, arithmetic operations, powers and simple roots

- **Equations and Inequalities:** linear and related equations and inequalities, simple second-degree equations, and analytic and non-analytic solution methods
- **Co-Ordinate Systems:** representation and description of data, position and relationships
- **Relationships Within and Among Geometrical Objects in Two and Three Dimensions:** static relationships such as algebraic connections among elements of figures (e.g. the Pythagorean theorem as defining the relationship between the lengths of the sides of a right triangle), relative position, similarity and congruence, and dynamic relationships involving transformation and motion of objects, as well as correspondences between two- and three-dimensional objects
- **Measurement:** quantification of features of and among shapes and objects, such as angle measures, distance, length, perimeter, circumference, area and volume
- **Numbers and Units:** concepts, representations of numbers and number systems (including converting between number systems), including properties of integer and rational numbers, as well as quantities and units referring to phenomena such as time, money, weight, temperature, distance, area and volume, and derived quantities and their numerical description
- **Arithmetic Operations:** the nature and properties of these operations and related notational conventions
- **Percentages, Ratios and Proportions:** numerical description of relative magnitude and the application of proportions and proportional reasoning to solve problems
- **Counting Principles:** simple combinations
- **Estimation:** purpose-driven approximation of quantities and numerical expressions, including significant digits and rounding
- **Data Collection, Representation and Interpretation:** nature, genesis and collection of various types of data, and the different ways to analyse, represent and interpret them
- **Data Variability and its Description:** concepts such as variability, distribution and central tendency of data sets, and ways to describe and interpret these in quantitative and graphical terms
- **Samples and Sampling:** concepts of sampling and sampling from data populations, including simple inferences based on properties of samples including accuracy and precision
- **Chance and Probability:** notion of random events, random variation and its representation, chance and frequency of events, and basic aspects of the concept of probability and conditional probability

### ***Mathematical Literacy – Contexts***

Context is the aspect of an individual's world in which a problem is set. All PISA mathematics items are set in a real-life context; however, that does not mean all the items are based on real-life events or scenarios. Some units are based on fictional but plausible scenarios where mathematics can be applied in various ways towards solving problems. The strategies used to solve a problem can be dependent on the context in which the problem is set, but care is taken to ensure that context-specific knowledge is not needed to solve any problem. The PISA 2022 mathematics assessment uses the same four contexts as in previous cycles, which are: personal, occupational, societal, and scientific. Note that there is no reporting by context but having these different classifications helped ensure that the items reflected a broad range of situations where mathematics could be encountered in real life. A brief description of each context follows.

- **Personal:** problems classified in the personal context category focus on activities of one's self, one's family or one's peer group. The kinds of contexts that may be considered personal include (but are not limited to) those involving food preparation, shopping, games, personal health, personal transportation, recreation, sports, travel, personal scheduling, and personal finance.



- **Occupational:** problems classified in the occupational context category are centred on the world of work. Items categorised as occupational may involve (but are not limited to) such things as measuring, costing and ordering materials for building, payroll/accounting, quality control, scheduling/inventory, design/architecture and job-related decision making either with or without appropriate technology. Occupational contexts may relate to any level of the workforce, from unskilled work to the highest levels of professional work, although items in the PISA survey must be accessible to 15-year-old students.
- **Societal:** problems classified in the societal context category focus on one's community (whether local, national, or global). They may involve (but are not limited to) such things as voting systems, public transport, government, public policies, demographics, advertising, health, entertainment, national statistics, and economics. Although individuals are involved in all of these things in a personal way, in the societal context category, the focus of problems is on the community perspective.
- **Scientific:** problems classified in the scientific category relate to the application of mathematics to the natural world and issues and topics related to science and technology. Particular contexts might include (but are not limited to) such areas as weather or climate, ecology, medicine, space science, genetics, measurement and the world of mathematics itself. Items that are intra-mathematical, where all the elements involved belong in the world of mathematics, fall within the scientific context.

## Role of the mathematics expert group in item development

As the contractor for mathematics instrument development, Core A was responsible for working with the Mathematics Expert Group (MEG) to understand their vision for the range and types of items to be developed for PISA 2022. To facilitate the transition from the work of Core B1 (framework development) to the instrument development activities, Core A retained the members of the MEG who met under Core B1 to begin work on the updated mathematics framework in 2017, and which continued into 2018.

Core A's work with the MEG began in February 2018 and focused on the following tasks:

- describing the kinds of items needed to assess the skills and abilities in the domain, particularly defining the behaviours associated with mathematical reasoning
- reviewing and understanding the proposed assessment design to determine the distribution of mathematics content across the major components of the framework
- defining the intersection between the kinds of functionality that might be desirable for measuring the construct and the functionality that was practical to implement in the assessment
- developing illustrative examples of tasks that reflect some of the new content and possible functionality of the platform.

Work with the subject matter experts continued beyond the initial meetings and went through instrument development and data analysis. For mathematics, MEG members reviewed assessment tasks as they were developed, provided input into the analysis of the field trial data, approved the set of items for the main survey, and worked with development and analysis staff to develop the described proficiency scales used for reporting the PISA 2022 results.

## PISA 2022 test development

Test development for the PISA 2022 cycle began in early 2018 and focused on the development of mathematics items for a computer-based assessment. For example, the following list from the updated

mathematics framework presents a few possible ways in which the computer platform was leveraged to assess mathematical literacy:

- Simulation in which a mathematical model has been established and students can change the variable values to explore the impact of the variables to create “an optimal solution”.
- Fitting a curve (by selecting a curve from a limited set of curves provided) to a data set or a geometric image to determine the “best fit” and using the resulting best fit curve to determine the answer to a question about the situation.
- Budgeting situations (e.g. online store) in which the student must select combinations of products to meet achieve a range of objectives within a given budget.
- Purchase simulation in which the student selects from different loan and associated repayment options to purchase an item using a loan and meeting a budget. The challenge in the problem is to understand how the variables interact.
- Problems that include visual coding to achieve a given sequence of actions.

However, it is important to note that not every new unit or item was developed requiring the use of some type of computer-based functionality. Item development efforts strove to maintain a balance between purposeful uses of the available technology, but the focus was always on assessing mathematical literacy and not information and communications technology (ICT) skills. To help with this last point, in addition to the general orientation, which provided students with an overview of the platform and standard functionality (e.g. navigating the interface, using drag-and-drop, selecting vs entering a response, etc.), item-specific tutorials and practice opportunities were built-in to every unit/item that used “novel” functionality (e.g. spreadsheets) before students could advance to the actual items. Even after students advanced past the requisite practice screens, instructions for using the specific tool in a unit were always available via drop-down menu at the top of each screen in the unit.

### ***Computer-based assessment: Screen design and interface***

The screen design and interface developed for the PISA 2015 cycle, and which was used for the PISA 2018 cycle, was again used for the PISA 2022 cycle.

#### *Navigation*

As in PISA 2015 and PISA 2018, students could navigate through the items as needed. For most units, students were able to move back and forth between items *within* a unit. They were not, however, able to move back and forth *between* units. Once students clicked on the “NEXT” arrow 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. There were a few exceptions regarding navigation *within* units where students were not permitted to return to a previous item. These within-unit restrictions were primarily used when information in a later item might help with answering an earlier item or in instances where it was desired that the students either have access or no longer have access to a tool. When students would click on the “Next” arrow a message would pop up indicating that it, “...will not be possible to return to this work.”, and students would have to click on “Yes” or “No” to indicate if they were ready to continue to the next question in the unit.

#### *Response modes*

Across all domains, PISA 2022 included items requiring one of five different response modes:

- **Selection items:** single-selection multiple choice; multiple-selection multiple choice (click on one or more options); complex multiple choice (table with statements and typically several yes/no or

true/false options); data selection (selecting rows of student-generated data to support or refute a claim); or click on an image

- **Numeric entry:** only numbers, commas, periods, dashes, and backslashes could be entered
- **Text entry:** a scrolling text box that did not constrain the length of a student response (consistent with what was possible for paper-and-pencil items); or certain mathematics items that used the equation editor
- **Drop-down menus**
- **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 mathematics section of the assessment, students were introduced to the calculator and the equation editor and given an opportunity to practice using each of these tools.

### ***Trend items***

The computer-based trend reading item pool contained 197 items (152 developed in PISA 2018 and 45 developed prior to PISA 2015), in addition to the 60 reading fluency items. Of the 197 trend reading items, 64 were human coded.

The computer-based trend science item pool contained 115 items (76 developed in PISA 2015 and 39 developed prior to PISA 2015) in six clusters. For science, these were the same trend clusters that were used in PISA 2018 and which remained intact for the PISA 2022 field trial and main survey. Of the 115 trend science items, 32 were human coded.

The computer-based trend mathematics item pool contained 74 items, 16 of which were human coded. The financial literacy item pool contained 46 items (five items developed in 2022 and 41 items developed prior to 2022). There were 16 human coded items in financial literacy.

For the “new” paper-based assessment there were: 66 science items (nine human coded), 66 reading items (37 human coded), and 62 mathematics items (40 human coded). For the one country taking the older paper-based assessment, there were: 85 science items (32 human coded), 87 reading items (51 human coded), and 71 mathematics items (38 human coded).

### ***New items***

For PISA 2022, test development occurred for the domains of mathematics, creative thinking, and financial literacy. To prepare for the implementation of the multistage adaptive design in the main survey, twelve 30-minute clusters of new items were developed for mathematics. In total, 61 new units with 182 new mathematics items were selected and included in the field trial. For financial literacy, three new units were developed with five total new items, all of which were retained for the main survey.

For information on the development of creative thinking, refer to chapter 4 of this technical report.

### ***International test development team***

Test development efforts for the mathematics assessment were coordinated by ETS as the Core A Contractor. As is the case with any large-scale international survey, it is important that the material used

in PISA reflect the range of contexts and experiences of students across participating countries/economies. One way to meet this goal was by convening an international team of item developers. For PISA 2022, the international test development team included individuals from the University of Luxembourg and the University of Liège. A second way to meet this goal was to work with countries/economies on development of materials. Core A provided countries/economies with a range of opportunities for participation during the development process.

### *National submissions*

The active involvement of countries/economies in the development process is important for the instruments to be internationally valid and representative. Thus, it was important to ensure that the final item pool reflected the international context of an assessment such as PISA. For example, Core A offered two item-development workshops, as well as accepted item submissions via the PISA Portal. This phase of the item-development process primarily occurred between April and September of 2018.

### *Item development workshops and submissions*

Two item development workshops were offered as part of the PISA 2022 efforts to involve countries/economies in the test development process. These took place in May and June of 2018 in Princeton, NJ, USA and in Liège, Belgium, respectively. Fifty-three participants from 29 countries/economies attended these workshops. From the test developers' point of view, the workshops made the development process more efficient because of the in-person training and collaboration, which was reflected in the quality of items that came out of the workshop and the items that were submitted subsequently. These workshops allowed representatives from countries/economies to interact and share ideas and expertise with members of the test development teams. Participants in the workshops wrote and reviewed items during the workshop and received some "real-time" feedback from the test development teams. The workshops also provided a venue to exchange ideas for ways to assess the content in the updated framework.

Overall, the item writing workshops and item submission process were extremely successful and ultimately resulted in 44 units with 130 new mathematics items that were used on the main survey. Additional new units were developed internally by experienced mathematics assessment specialists at ETS.

### *Item Reviews*

Newly developed units were submitted for translatability review at the same time they were released for country/economy review. 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 PISA 2022 cycle alerted item developers to both general wording patterns and specific item wording that was known to be problematic for some translations and suggested alternative wordings. This provided item developers with the opportunity to make wording revisions at an early stage. In some cases, revisions were performed by simply using the alternatives provided and in others by working with cApStAn to explore a suitable wording that would lend itself to being translated without compromising what was being assessed.

All newly developed mathematics and financial literacy items were released for country/economy review prior to the field trial. Countries/economies had two weeks to perform reviews and submit feedback on all draft items. Mathematics items were released in four batches between September and December 2018. Test developers received review forms from 40 countries/economies for Batch 1, 54 countries/economies for Batch 2, 53 countries/economies for Batch 3, and 54 countries/economies for Batch 4. The newly developed financial literacy items were released in one batch, which was reviewed by 19 countries/economies.

Preparation of the French source version for all new mathematics units provided another opportunity to identify issues with the English source version related to content and expression. 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, cognitive labs were conducted by the University of Luxembourg and by the University of Liège. A total of 11 new mathematics units (five at the University of Luxembourg and six at the University of Liège) were evaluated as part of these cognitive labs. The 11 units contained a mixture of new content and/or new functionality. These cognitive labs provided useful information to test developers concerning students' understanding of the content and what the items were assessing, response formats, the clarity of instructions and introductions, how the interactive elements functioned, and timing. The results led to improvements in the 11 items used in the cognitive labs, as well as provided test developers with some general guidelines to apply to all new units.

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

The PISA 2022 item development process produced a total of 61 new mathematics units with 182 items that were selected for use in the field trial. Items were selected for inclusion in the field trial based on country/economy reviews, feedback from the mathematics expert group and the distribution of items across the key categories as defined in the framework. Of these 182 new mathematics items, 74% were submitted by participating countries/economies (from the item development workshops and item submissions via the Portal), and 26% were developed by ETS's test development team.

## **Field trial**

The PISA 2022 field trial data collection timeline began in March 2020 but was quickly disrupted by the COVID-19 global pandemic. Even though 17 participants were able to complete a limited field trial in 2020, most participants postponed the field trial until 2021. Of the 17 participants that administered the limited field trial in 2020, six participants chose to readminister the field trial in 2021. In total, 83 countries/economies (79 that administered on computer and four that administered on paper), consisting of 142 language versions, participated in this cycle of PISA. 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 2022, the Core A contractors continued to assume responsibility for assembling the assessment instruments for both paper- and computer-based countries/economies. Countries/economies were responsible for translating all new material and performing both linguistic and layout quality control checks for trend and new items.

### *Computer-based trend items*

Countries/economies that participated in the PISA 2015 and/or PISA 2018 computer-based assessment, were given access to the existing XLIFFs (XML Localization Interchange File Format) files from the previous administration and had the opportunity to review their materials for any errors or necessary updates.

For countries/economies switching from a paper- to a computer-based assessment, the Core A contractors copied their material into the computer-readable XLIFF that was used for the computer-based instruments. This was done both as a quality control process and to reduce the number of tasks assigned to countries

given the short development timeline. Once the XLIFF files were created, the Core A contractors asked the countries/economies to perform a review comparing the new computer versions against PDF files of their paper-based items.

In both cases, countries/economies were asked to document any errors, which included typographical mistakes or text errors introduced in the process of copying and pasting across formats. All 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 corrected by the Core A technical team. As an additional quality control check, the Core A contractor also performed layout checks of all items in all languages to identify errors that may have been missed.

### *Computer-based new items*

All new mathematics and financial literacy items needed to be translated following the translation and reconciliation processes defined in the PISA standards. Following verification of the translations and the correction of any remaining errors, countries/economies were asked to sign off on their cognitive materials and those files were then considered locked for use.

### *Preparing the field trial national student delivery systems (SDS)*

The Student Delivery System (or SDS) was again used for PISA 2022 and was a self-contained set of applications for delivery of the computer-based cognitive assessments and computer-based student background questionnaires. A master version was assembled first for countries to test within their national IT structure. This allowed countries/economies to become familiar with the operation of the SDS and to check the compatibility of the software with the computers being used to administer the assessment.

Once all the cognitive and background materials were approved and locked, the SDS was assembled and tested first by the Core A technical team. The SDS was then released for national testing. Countries/economies were asked to check their SDS following a specific testing plan provided by Core A and to identify any residual content or layout issues. If issues were identified, they were corrected by the Core A technical team, and a second SDS was released. Once countries/economies signed off on their SDS, their instruments were released for the field trial.

### *Paper-based instruments*

National versions of the paper-based trend clusters were again prepared by the Core A contractor. To better ensure comparability of the paper-based assessment materials across countries/economies and languages, digital files of the booklets were centrally created by Core A and then reviewed and approved by countries/economies. Those countries/economies who were new to PISA needed to translate those materials following the standard translation and verification process. Existing paper-based countries/economies needed to update the common booklet parts (which included the cover, general instructions, formula sheet for mathematics, and the acknowledgements page), while new countries/economies had to translate these materials.

The approved clusters were then assembled into field trial paper booklets by the contractors using a centralised process that ensured comparability of layout. As a final step, the assembled booklet files were released and participants performed a final review and Core A implemented any changes, as needed. This process continued until National Centres had approved, print-ready files.

### ***Field trial coding***

Coding guides for trend items were compiled by Core A based on previous national versions. Continuing a practice that started in the PISA 2018 cycle, separate guides were updated/prepared for computer-based and paper-based participants.

The English master versions of the new mathematics and new financial literacy coding guides were released in draft form prior to the coder training meeting in January 2020. Based on discussions at that meeting, the coding guides were finalised and the updated English versions, along with the French source version (for new mathematics), were released to countries/economies in March 2020, prior to the beginning of the field trial data collection period. For the trend domains, a similar process was followed but with corrections to the guides restricted to correcting outright mistakes or providing some additional examples for clarification purposes.

#### *Field trial coder training*

The international field trial coder training was held in-person in January 2020 with sessions for all domains, including separate sessions for paper-based participants. The goals of the training included having attendees (master coders) develop an in-depth understanding of the coding rules for each item, so they would be prepared to train coders in their countries/economies and reaching consensus about the coding rules to better ensure consistency of coding both within and between countries/economies and across cycles. Trainers reviewed the content 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. When there were disagreements about coding for an item, they were discussed so that all attendees understood the specific coding rules for that item.

Due to the postponement of most field trials in 2020, field trial coder trainings were held virtually in January and February of 2021 for new mathematics, creative thinking, and financial literacy (new items only). The virtual training also included a recap of general coding principles and procedures, as well as a refresher training on the open-ended item coding system (the OECS).

#### *Field trial coder queries*

As was the case during previous cycles, Core A set up a coder query service for the PISA 2022 field trial. Countries/economies were encouraged to send queries to the service so that a common adjudication process was consistently applied to all coder questions about human-coded items. Queries were reviewed, and responses were provided by domain-specific teams that included item developers, and for trend items, members of the response team from previous cycles. For the new items, the coder query service was particularly valuable as it provided item developers with a better sense of the “range” of responses that could be expected, which in turn led to refinements of the coding guide.

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

### ***Field trial outcomes***

The PISA 2022 field trial was designed to yield information about the quantity and quality of data collected as well as to prepare the multistage adaptive testing design for the main survey. More specifically, general 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 functioning 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 in both the 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. Furthermore, the field trial results confirmed the feasibility of introducing multistage adaptive testing in the main survey as unit order effects were found to be negligible.

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 2021. That batch included data from 52 countries/economies, of which 41 carried out the field trial in 2021 and 11 in 2020. Of those, one participant administered the paper-based assessment, 51 administered the computer-based assessment, and one conducted data collection in 2020 and in 2021. The field trial analyses were updated after receiving additional data, which increased the number of participating countries/economies to 80 by the end of 2021. Of these, three participants implemented the field trial as a paper-based survey and 77 that implemented it as a computer-based survey.

## Main survey

The PISA 2022 main survey was conducted between March and December 2022. Most countries/economies completed the main survey data collection by May 2022. 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 with the items still in need of correction. The Core A contractors worked with countries/economies 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 September 2021 and was conducted on a rolling basis based on main survey start dates.

Following release of the field trial data, countries/economies completed 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 and participants were asked to provide comments about these specific items where they could identify serious errors. Requests for corrections were reviewed by Core A, and if approved, implemented.

### ***Item selection for Mathematics***

The initial selection of mathematics items for the main survey was a collaborative effort between the test development team and psychometricians based mostly on item statistics from the first batch of field trial data. The first step was to generate a list of flagged items based on the following statistics and associated criteria:

- Scoring reliability rater agreement (below 0.92%)
- Percentage of omitted responses (above 20% in each country/economy)
- IRT discrimination and difficulty parameters ( $a < 0.1$  or  $|b| > 5$ )
- IRT MD and RMSD fit statistics (0.15 for new items and 0.20 for trend items)
- Item-level and unit-level response time (more than three minutes per item)



Next, the list of flagged items was reviewed from a content perspective with an aim towards removing any items with possible content flaws or items that were not able to be scaled appropriately. Another factor influencing main survey item selection was feedback from National Centres. Participants were asked to rate each item from the field trial with regards to how common the content was to their national curriculum using the following values: 1 = not in curriculum, 2 = in some curriculum, or 3 = standard curriculum material. They were also asked to rate each item on how relevant each item was to “preparedness for life” using the following values: 1 = not relevant, 2 = somewhat relevant, or 3 = highly relevant. The final step was a review of the remaining items, based on the degree to which they had been flagged (i.e. items that had stronger statistics were kept over those with weaker statistics), but also to determine if removing certain items would lead to an imbalance in domain representation (according to the target construct distributions in the framework), and to check for any changes to how a unit would function if an item or items were removed (e.g. if an item was removed that introduced or built on the scenario which the unit was written about, so that a subsequent item became unclear because it referenced information no longer present in the unit).

Once this review process was completed, a total of 30 mathematics items (22 new items and eight trend items) were dropped from across 20 units (15 new units and five trend units). A total of seven units (five new units and two trend units), which consisted of 17 items (12 new items and five trend items), were dropped completely. The remaining dropped items came from units where one or more items were retained for the PISA 2022 main survey. The resulting computer-based mathematics item pool for the main survey contained 99 total units (56 new units and 43 trend units) and 234 total items (160 new items and 74 trend items). For the paper-based designs, no items or units were dropped following the field trial.

### ***Assigning mathematics units to the multistage adaptive design***

The multistage adaptive design for mathematics expanded and enhanced what was accomplished with the adaptive design for reading in PISA 2018. Test assembly for PISA 2022 was implemented in four steps:

1. Assemble non-overlapping parallel item sets.
2. Assemble core and adaptive testlets from each item set.
3. Assemble multistage paths using the core and adaptive testlets.
4. Assemble linear forms using the core testlets.

Also, for PISA 2022 automated test assembly (ATA) was employed to assemble the test paths and forms via mixed-integer linear programming. This was done using commercial software. The software provided a principled design approach and was able to much more efficiently handle the large number of decision variables and constraints at each step of the assembly process. Note that there was some flexibility with constraints when creating the core and adaptive testlets as long as all constraints were met in the full path or form. A summary of some key features – framework distributions and psychometric properties – of the four steps follows.

#### *Non-overlapping parallel item sets*

Each of the three item sets contained 78 items and 33 units. Each unit only appeared in one item set. The maximum score of each set was either 99 or 100 points. Each set contained approximately 27% trend items. Approximately 85% of the items in each set were machine coded, and across all sets there were approximately equal numbers of items for each of the four major item types used in PISA (simple multiple choice, complex multiple choice, computer-scored open response, and human-coded open response). Each set contained approximately 24% of items from change and relationships, 32% from quantity, 18% from space and shape, and 26% from uncertainty and data. Each set contained approximately 32% employ items, 21% formulate items, 24% interpret/evaluate items, and 23% reasoning items.

### *Core and adaptive testlets from each item set*

Each of the core testlets in the three item sets contained from three to five, three to six, or four to five units, and nine to 10 total items. The maximum score per core testlet, across all items sets, was from 12 to 14 points, of which human-coded items contributed from two to four points (the number of human-coded items in each core testlet ranged from one to two or one to three across all item sets). The maximum number of common items was set at six, so the percent overlap was either 27% or 28% depending on the item set. Percent overlap is the number of test pairs with overlap divided by the total number of test pairs. The core testlets had a percent connectedness of either 20% or 21%, depending on the item set. Percent connectedness is the number of unit pairs in tests divided by the total number of unit pairs. The median total response times for the core testlets ranged between 11 and 13 minutes across all item sets.

Each of the stage 1 adaptive testlets in the three item sets contained from three to five or three to six units, and nine to 10 total items. The maximum score per stage 1 testlet, across all items sets, was from 12 to 14 points, of which human-coded items contributed from two to three or two to four points (the number of human-coded items in each stage 1 testlet ranged from one to two or zero to three across all item sets). The percent overlap ranged from 25% to 27%, depending on the item set. The stage 1 testlets also had a percent connectedness of either 20% or 21%, depending on the item set. The median total response times for the stage 1 testlets also ranged between 11 and 13 minutes across all three item sets.

Each of the stage 2 adaptive testlets in the three item sets contained from three to five or from three to six units, and nine to 10 total items. The maximum score per stage 2 testlet, across all items sets, was from 12-13 or 11-14 points, of which human-coded items contributed from one to two, two to three, or zero to five points (the number of human-coded items in each stage 2 testlet ranged from one to two, one to three, or zero to three across all item sets). The percent overlap ranged from 23% to 26%, depending on the item set. The stage 2 testlets had a percent connectedness of either 19% or 20%, depending on the item set. The median total response times for the stage 2 testlets again ranged between 11 and 13 minutes across all item sets.

### *Multistage paths using the core and adaptive testlets*

A total of 192 adaptive paths in the mathematics assessment were implemented for the PISA 2022 main survey. The number of units per path ranged from 10 to 16 with a median of 13 units. The number of items per path ranged from 28 to 30 with a median of 30 items. The number of trend mathematics items ranged from 3 to 16 with a median of 9, while the number of new mathematics items ranged from 14 to 27 with a median of 20. The median number of items by content area for each path was seven for change and relationships, 10 for quantity, five for space and shape, and seven for uncertainty and data. The median number of items by process for each path was nine for employ, six for formulate, seven for interpret/evaluate, and seven for reasoning. For both the content areas and the process classifications, the percentage distributions in each testlet mirrored the distributions of the entire mathematics item pool. Each unit appeared on average in 24.5 paths. The overlap percentage across all 192 paths was 75% (i.e. 75% of the possible pairs of paths have at least one unit in common). The percentage of observed unit pairs was 78% (i.e. 78% of the possible pairs of units were observed). For comparison, in PISA 2018, the percentage of observed unit pairs in the reading MSAT design was only 55%.

### *Linear forms using the core testlets*

A total of 48 linear forms were in the PISA 2022 main survey mathematics assessment. The linear forms were comprised of the 48 core testlets. The number of units per form ranged from 11 to 15 with a median of 13 units. The number of items per form ranged from 29 to 30 with a median of 30 items. The number of trend mathematics items ranged from 1 to 19 with a median of 10, while the number of new mathematics items ranged from 11 to 29 with a median of 20. The median number of items by content area for each

form was six for change and relationships, 10 for quantity, five for space and shape, and eight for uncertainty and data. The median number of items by process for each form was nine for employ, five for formulate, seven for interpret/evaluate, and seven for reasoning. For both the content areas and the process classifications, the percentage distributions in each linear form mirrored the distributions of the entire mathematics item pool.

After the four steps above were completed by the psychometrics team, all the proposed testlets were reviewed by the mathematics development team to look for any potentially problematic unit pairings (e.g. having multiple units within a testlet that assess the same construct) and to propose recommended changes. The development team then worked closely with the psychometricians to determine the effect the proposed changes would have on the design, and to make additional changes if needed. Once the unit pairings in each testlet were finalized, the development team made recommendations for how to order the units within each testlet.

### ***Review by the Mathematics Expert Group***

Once the item selection was complete and the units were assigned to the multistage adaptive design, Core A psychometricians performed simulation studies to assess the performance of the design using the preliminary item parameters obtained from the field trial. The details of these simulation studies are described in Yamamoto, Shin and Khorramdel (2018<sub>[1]</sub>). In short, the simulation studies suggested that the item parameters could be recovered well with minimal errors and that the proposed multistage adaptive design would improve the measurement precision for all ranges of skill distribution, particularly at the lower and higher ends of distribution. Specifically, the simulation study showed a gain in measurement precision of 10.6% at the lowest proficiency level, and a 13% gain at the highest proficiency level.

Given that the multistage adaptive testing design consisted of 192 possible paths, it was not possible for the mathematics experts to review all those combination of item sets and make recommendations for the selection. Instead, at the MEG meeting following the field trial, a thorough explanation of the item selection process and the characteristics of the main survey item pool were presented and discussed. The item pool was evaluated at a holistic level, considering the representation of the content areas and cognitive processes across the entire pool, including the distributions of difficulty and construct representation within each stage of the multistage adaptive design. At the end of the meeting, the experts signed off on the main survey item pool and the multistage adaptive design.

### ***Construct coverage***

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

A total of 234 items – 74 trend and 160 new items – were selected for the computer-based mathematics assessment, and those 234 items represent a total of 253 possible score points. Annex Table 3.A.6 shows the item counts, score points and percentage of score points by cognitive process and by content area for the main survey CBA mathematics items.

Of the 160 new items retained for the main survey, 74% were originally submitted by countries/economies (from either the item development workshops or item submissions) and 26% were created by test developers at ETS.

### ***Financial Literacy***

Item selection for financial literacy was based on classical item analyses. All five new items were retained for the main survey and two trend items – one from each cluster – were recommended by the PISA Technical Advisory Group (TAG) to be dropped based on concerns over the amount of time that students

were spending on those two items. A total of 46 items (41 trend and 5 new) were used in the main survey financial literacy assessment. Annex Table 3.A.7. shows the distributions of the 46 financial literacy items across the two aspects of the framework: process and content.

The paper-based and computer-based item counts for reading, mathematics, science, creative thinking, and financial literacy in both the field trial and main survey are presented in Annex Table 3.A.8.

## **Preparation of data collection instruments**

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

The process for creating the main survey student delivery system (SDS) followed the approach 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 the materials were agreed upon, they were digitally locked, and it was not possible to edit or change them. This included the questionnaires and cognitive instruments. The student delivery system was then assembled and tested first by Core A. Countries/economies were then asked to check their SDS and identify any remaining content or layout issues. Once countries/economies signed off on their national SDS, their final systems were released for the main survey.

### *Preparing main survey paper-based instruments*

As in the field trial, national versions of the main survey paper-based booklets were centrally prepared by the Core A contractor to better ensure comparability of the paper-based assessment materials across participants and languages. Once the workflow for reviewing field trial data and requesting changes to items was completed, and the common booklet parts (i.e. cover page, formula sheet, general instructions) were updated as needed, the approved materials were assembled into main survey booklet files by Core A. The booklet files were then sent to the countries/economies for review. If any changes were needed, Core A would implement them, and the process for reviewing the files would repeat until the National Centre approved all files for printing.

## **Main survey coding**

Coder training for the main survey was conducted virtually for all domains. For mathematics and creative thinking, full trainings were offered for all main survey items (trend and new). The trainings for reading and science were targeted on items that were typically more challenging to code (e.g. items with low reliability rates or items with a high number of coder queries). The training for financial literacy covered all the new items but was targeted for the trend items, using the same criteria that reading and science used to identify 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 particularly challenging responses. Queries were reviewed, and responses were provided by domain-specific teams including item developers and members of the response team from previous cycles. Revisions were made to the coding guides for mathematics and creative thinking, and to the new financial literacy items following the field trial. The coder queries helped test developers see response categories that were not anticipated during the initial development of the coding guide. Thus, based on the queries received, test developers made some coding guides clearer and added sample responses to the guides to better illustrate the range of, and different types of, responses. Workshop examples were also enhanced by adding more authentic student responses that better illustrated the boundaries between full credit, partial credit (if applicable) and no credit. Following the international coder trainings, additional revisions were made to the mathematics,

creative thinking, and financial literacy (new items only) coding guides in response to discussions that took place during the trainings.

### ***Released items to illustrate the framework***

As has been the case in previous PISA cycles, several items were released to the public domain at the time of publication of the PISA 2022 results to illustrate the kinds of items included in the assessment. Following the field trial, a list of proposed units to release was reviewed by the MEG and the OECD, and after the main survey, another list of proposed units to release was reviewed by the MEG and the OECD. The following four new mathematics units were approved for release after the field trial: *Car Purchase* (2 items), *DVD Sales* (3 items), *Moving Truck* (2 items), and *Spinners* (3 items). After the main survey, the following four new mathematics units were approved for release: *Solar System* (2 items), *Triangular Pattern* (3 items), *Points* (1 item), and *Forested Area* (4 items). These units are available at [www.oecd.org/pisa](http://www.oecd.org/pisa).

## References

- Yamamoto, K., H. Shin and L. Khorramdel (2018), “Multistage adaptive testing design in international large-scale assessments”, *Educational Measurement: Issues and Practice*, Vol. 37/4, pp. 16–27. [1]
- 

## Note

1. For a complete description of the PISA 2022 Mathematics Framework, please visit the site <https://pisa2022-maths.oecd.org>.

## Annex 3.A. Test developments for the core domain

Annex Table 3.A.1. Chapter 3: Test developments

Tables	Title
Table 3.A.2	Number of PISA items by core domain and across cycles in the main survey
Table 3.A.3	Domain coverage for PISA 2022: CBA
Table 3.A.4	Main survey domain coverage for PISA 2022: PBA
Table 3.A.5	Main survey computer-based UH form design
Table 3.A.6	Item counts and score points of the main survey CBA mathematics items by framework categories
Table 3.A.7	Main survey financial literacy item counts by framework categories
Table 3.A.8	Item counts in the field trial and main survey by domain and delivery mode

Annex Table 3.A.2. Number of PISA items by core domain and across cycles in the main survey

	2000	2003	2006	2009	2012	2015	2018	2022
Reading	129	28	28	131	44	103	245	197
Mathematics	43	84	48	35	109	83	83	234
Science	45	34	103	53	53	184	115	115

Note: Red font colour = Major domain for that cycle.

For the 2015 and 2018 cycles, the computer-based mathematics instrument contained 82 items, while the equivalent paper-based instrument contained 83 items. This is because there was one item that was not able to be transitioned to a computer-based delivery in 2015 (the item requires students to draw on a map).

The number of mathematics items in the 2022 cycle includes 74 "trend" items (i.e. items developed prior to this cycle) and 160 "new" items (i.e. items developed this cycle).

Annex Table 3.A.3. Domain coverage for PISA 2022: CBA

Domain	Field trial		Main survey		Total items – 2022 MS
	New	Trend	New	Trend	
Reading Literacy	No new item development for 2022	Adaptive design: 197 items	No new item development for 2022	Same as Field Trial Trend	197
Scientific Literacy	No new item development for 2022	6 clusters: 115 items (76 from the 2015 cycle; 39 used prior to 2015)	No new item development for 2022	Same as Field Trial Trend	115
Mathematical Literacy	12 clusters: 182 items	7* clusters: 82 items	Adaptive design: 160 items	Adaptive design: 74 items	234
Creative Thinking	5 clusters: 38 items	New domain – no trend items	5 clusters: 36 items	New domain – no trend items	36
Financial Literacy	3** units: 5 items	2 clusters: 43 items	5 items	41 items	46

Note: Each cluster was designed to take approximately 30 minutes of testing time.

\* For the PISA 2022 cycle field trial, there were actually 7 trend mathematics clusters because all computer-based participants administered the units from clusters M6a ("standard items") and M6b ("easier items"). In previous administrations, participants administered either M6a or M6b but not both.

\*\* There are two financial literacy clusters - F1 and F2 - used in both the field trial and main survey this cycle. However, only 3 new units ( 5 total items) were developed for this cycle, and they were distributed across the two existing clusters (two new units in cluster F1 and one new unit in cluster F2).

**Annex Table 3.A.4. Main survey domain coverage for PISA 2022**

PBA Instrument Used by One Participant this Cycle	
Domain	Field trial and main survey
Reading	6 clusters: 87 items Same set of items that all PBA participants used in 2018 and 2015 Prior to 2015, these items were last used in 2012 and 2009
Science	6 clusters: 85 items Same set of items that all PBA participants used in 2018 and 2015 Prior to 2015, these items were last used in 2012, 2006 and 2003
Mathematics	6 clusters: 71 items Same set of items that PBA participants used in 2018 and 2015 These items were all taken from the 2012 cycle
New Instrument Used by All Other PBA Participants this Cycle	
Domain	Field trial and main survey
Reading	4 clusters: 66 items*
Science	4 clusters: 66 items
Mathematics	4 clusters: 63 items*

Note: \* There are 64 items in the new PBA mathematics assessment; however, one of the items is actually a reading item (it is in a set that contains a mathematics and a reading item), so there are only 63 items that contribute towards the mathematics scale.

**Annex Table 3.A.5. Main survey computer-based UH form design**

Form	Cluster 1	Cluster 2	Cluster 3	Cluster 4
99 (UH)	MU1	MU2	RU1	SU1

Note: Where M = mathematics, R = reading, and S = science.

**Annex Table 3.A.6. Item counts and score points of the main survey CBA mathematics items by framework categories**

	Trend Items	New Items	Combined (Trend + New)	Dichotomously Scored Items (1 point each)	Polytomously Scored Items (2 points each)	Total Score Points*		Framework Recommendation
<b>Cognitive process</b>	<b>Count</b>	<b>Count</b>	<b>Count</b>	<b>Count</b>	<b>Count</b>	<b>Points</b>	<b>%</b>	<b>%</b>
Formulating situations mathematically	11	37	48	47	1	49	19%	25%
Employing mathematical concepts, facts and procedures	24	51	75	72	3	78	31%	25%
Interpreting, applying and evaluating mathematical outcomes	10	47	57	55	2	59	23%	25%
Reasoning	29	25	54	41	13	67	26%	25%
Total	74	160	234	215	19	253	100%	100%
<b>Content area</b>	<b>Count</b>	<b>Count</b>	<b>Count</b>	<b>Count</b>	<b>Count</b>	<b>Points</b>	<b>%</b>	<b>%</b>

Change and relationships	17	38	55	50	5	60	24%	25%
Space and shape	17	26	43	39	4	47	19%	25%
Quantity	21	55	76	71	5	81	32%	25%
Uncertainty and data	19	41	60	55	5	65	26%	25%
<b>Total</b>	<b>74</b>	<b>160</b>	<b>234</b>	<b>215</b>	<b>19</b>	<b>253</b>	<b>100%</b>	<b>100%</b>

Note: \*The total score points are based on one point for each dichotomously scored item and two points for each polytomously scored item.

### Annex Table 3.A.7. Main survey financial literacy item counts by framework categories

Process	Number		Framework Recommendation
	Number	%	%
Identify financial information	7	15%	15-25%
Analyse information in a financial context	14	30%	15-25%
Evaluate financial issues	15	33%	25-35%
Apply financial knowledge and understanding	10	22%	25-35%
<b>Total</b>	<b>46</b>	<b>100%</b>	<b>100%</b>
Content	Number		%
	Number	%	%
Money and transactions	11	24%	30-40%
Planning and managing finances	16	35%	25-35%
Risk and reward	12	26%	15-25%
Financial landscape	7	15%	10-20%
<b>Total</b>	<b>46</b>	<b>100%</b>	<b>100%</b>

### Annex Table 3.A.8. Item counts in the field trial and main survey by domain and delivery mode

Domain	Field trial		Main survey	
	Paper-based (Design 1 / Design 2)	Computer-based	Paper-based (Design 1 / Design 2)	Computer-based
Reading	(87 / 66)	197 (+ 65 fluency items)	(87 / 66)	197 (+ 65 fluency items)
Mathematics	(71 / 63)	264	(71 / 63)	234
Science	(85 / 66)	115	(85 / 66)	115
Creative thinking	NA	38	NA	36
Financial literacy	NA	48	NA	46



# 4 Creative Thinking Test Design and Test Development

## Introduction

This chapter describes the assessment design for the PISA 2022 Innovative Domain: Creative Thinking (CT) as well as the processes used by the PISA Core B3 contractors, ACT and Cito, and the international test development team to develop the innovative domain assessment for the PISA 2022 cycle.

Activities for the innovative domain test design and test development included the following:

- The creation of a Creative Thinking Expert Group to guide test design and test development
- Development of a creative thinking assessment framework
- Assessment development
- Creative thinking validation studies
- Field Trial
- Main Survey

## The Role of the Creative Thinking Expert Group in Item Development

As the Core B3 contractor in charge of Creative Thinking instrument development, ACT was responsible for working with the creative thinking expert group (CTEG) as applicable. Work focused on understanding the CTEG's vision for the Creative Thinking framework as well as the range and types of items to be developed for PISA 2022 Creative Thinking assessment. CTEG members began work on the framework in September 2017 and finalized the framework September 2022. Core B3's work with the CTEG began in February 2018 and focused on the following tasks:

- describing the kinds of items needed to assess the skills and abilities in each domain as defined in the framework (OECD, 2019<sub>[1]</sub>).
- 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 testing functionalities (e.g. drawing tool, simulation, innovative item types) that would be desirable to develop for measuring the construct and would be feasible to implement in the context of PISA.

Work with the CTEG continued beyond the initial meeting through instrument development and data analysis. CTEG members played an important role in reviewing assessment tasks as they were developed, providing input into the analysis of the Field Trial (FT) data, approving the set of items for the Main Survey, and working with development and analysis staff to develop the described scales and performance level descriptors used for reporting the PISA 2022 Creative Thinking results.

## PISA 2022 Creative Thinking Assessment Framework

The PISA 2022 Creative Thinking assessment focused on the creative thinking processes that one can reasonably expect from 15-year-old students. It does not aim to single out exceptionally creative individuals, but rather to describe the extent to which students are capable of thinking creatively when searching for and expressing ideas, and how this capacity is related to teaching approaches, school activities, and other features of education systems.

The main objective of PISA is to provide internationally comparable data on students' creative thinking competence that have clear implications for education policies and pedagogies. The creative thinking processes in question therefore need to be malleable through education; the different enablers of these thinking processes in the classroom context need to be clearly identified and related to performance in the assessment; the content domains covered in the assessment need to be closely related to subjects taught in common compulsory schooling; and the test tasks should resemble real activities in which students engage, both inside and outside of their classroom, so that the test has some predictive validity of creative achievement and progress in school and beyond.

As the innovative domain for the PISA 2022 cycle, the creative thinking assessment focused on the skills that twenty-first century students need as organizations and societies around the world increasingly depend on innovation and knowledge creation to address emerging challenges, giving urgency to innovation and creative thinking as collective enterprises. The domain is defined as follows:

The competence to engage productively in the generation, evaluation, and improvement of ideas, that can result in original and effective solutions, advances in knowledge, and impactful expressions of imagination (OECD, 2019<sub>[1]</sub>).

Three cognitive facets that support creative idea generation and evaluation were further defined and included:

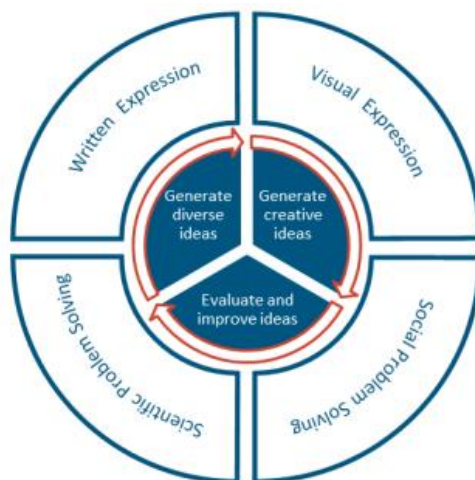
- **Generate Diverse Ideas (GDI):** students are asked to develop two or three ideas and are assessed on the appropriateness of these ideas (their alignment with the task requirements) as well as whether the two or three ideas are sufficiently different from one another.
- **Generate Creative Ideas (GCI):** students are asked to provide creative ideas and are assessed on the appropriateness of these ideas as well as whether the ideas occur with thematic infrequency.
- **Evaluate & Improve Ideas (EII):** students are asked to improve on the creativity of an idea that is provided to them and are assessed on whether the idea occurs with thematic infrequency.

As creative thinking can be expressed in a large number of possible applications, and the nature of these applications influence the knowledge and skills that are required to produce a creative output four domains were chosen for the PISA 2022 Creative Thinking assessment:

- Written Expression
- Visual Expression
- Social Problem Solving
- Scientific Problem Solving

The resulting competency model allows students the opportunity to demonstrate their capacity to generate, evaluate, and improve ideas across four distinct domains of applications. This design is expected to provide information about students' strengths and weaknesses across countries.

Figure 4.1. Competency model for the PISA test of creative thinking



Items were distributed across facets and domains to allow for a range of opportunities for expression. The distribution for the field trial included 14 generate diverse ideas items, 12 generate creative ideas items, and 12 evaluate & improve items. These are shown in Annex Table 4.A.2.

### PISA 2022 Innovative Domain Assessment Design

According to the assessment design, about 28% of the sample of PISA students were administered the creative thinking assessment. Students who took the creative thinking assessment spent one hour on creative thinking items with the remaining hour assigned to one of the other core domains (mathematics, reading, or science).

Creative thinking items were organized into test units. The units vary in terms of the facets that are measured, the domain and duration. Items were distributed within the units with some units having a single item and some units having multiple items.

Dependencies between items within units was minimized. The duration of each unit was between 5 and 15 minutes. The units were then organized into five mutually exclusive 30-minute blocks or clusters. The clusters were rotated according to the integrated design presented in Chapter 2 of this Technical Report. The assessment aimed to achieve a good balance between units that situate creative thinking and the four domains.

The items used to assess the creative thinking facets required of three different types of responses. Constructed-response tasks accounted for 92% of the items in the assessment. These typically call for a written response, ranging from a few words (e.g. cartoon caption or scientific hypothesis) to a short text (e.g. creative ending to a story or explanation of a design idea). Some constructed-response items call for a visual response (e.g. designing a poster combining a set of given shapes and stamps) that is supported by a simple drawing editor tool. The assessment also included two items that were part of an interactive simulation-based task which employs an interactive simulation environment and two items that consist of a task which calls for answers that are based on the choice of selecting a previously suggested idea or generating a new idea.

## PISA 2022 Innovative Domain Assessment Development

Test development for the PISA 2022 Creative Thinking assessment cycle began in early-2018 and focused on the development of items for a computer-based assessment. Through a process that included both CTEG contributions as well as country submission and country review, Core B3 along with the OECD selected a final set of item scenarios. Core B3 test developers further developed the scenarios. The OECD reviewed all scenarios and items early in the review process, prior to country reviews to ensure the items fulfilled the goals of the revised framework.

Newly developed units were submitted for translatability review at the same time they were released for country review. 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 PISA 2022 cycle alerted test developers to both general wording patterns and specific item wording that are known to be problematic for some translations and suggested alternatives. This allowed test 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.

To ensure that the creative thinking assessment items were understood the same way across linguistic and cultural groups, participating countries engaged in several cycles of review of the test material to help identify items that may be likely to suffer from cross-cultural bias. This enabled problematic cultural and linguistic characteristics to be identified during the early stages of the assessment development process. Countries had two weeks to perform reviews and submit feedback on all draft stimuli and items.

Preparation of the French source version for all new units provided another opportunity to identify issues with the English source version related to content and expression. Development of the two source versions helped identify 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.

Experienced testing professionals were engaged to conduct cognitive laboratory exercises with students in Australia, Singapore, and the United States. In the format of thinking-out-loud exercises, students around the age of the PISA population were asked to explain their thought processes in answering, and point out any difficulties or misunderstandings in the instructions or stimulus material. Information from these sessions was used to identify opportunities for revision and optimization of items as well as to correct several identified bugs (ACT, 2018<sup>[2]</sup>).

Validation Studies were conducted in parallel to the overall test development process, in an iterative manner, in order to observe how the then-current test materials functioned under similar test conditions. The purpose of each Validation Study was to provide evidence on the performance of creative thinking assessment in PISA-like classroom settings, collect sample student responses in multiple countries, assess the inter-rater reliability of human coded items (i.e. agreement between raters); determine the extent to which a creative thinking score or sub-scores can be obtained from the creative thinking assessment; and gain preliminary insights on the essential training materials needed for human coders.

A total of 703 15-year-old students from Singapore (206), Australia (234), and Canada (263) participated in the Validation Study between October to November 2018. Samples were recruited through PISA National Project Managers and coordinated with the OECD Secretariat.

The Validation Study instrument included 12 fully functional prototype units delivered in three forms, four units per form. Each form contained one unit per domain. Each unit included between 4-6 items (tasks). An analysis of the genuine student data indicated items that did not perform as intended (e.g. inter-rater scoring agreement, item difficulty, credit distribution), and informed evidence-based improvements to the test material, as well as development of and improvements to coder training material such as the coding

guide (ACT, 2019<sup>[3]</sup>). The validation study also helped refine the methodology followed for scoring students' responses and provided genuine responses for the international coder workshops.

## Field Trial

The Field Trial for creative thinking was initially scheduled for 2020; however, this timeline was disrupted by the COVID-19 pandemic, with findings to be further investigated during a second administration of the Field Trial in 2021. The limited field trial (LFT) conducted in 2020 with 11 countries provided preliminary evidence in support of: (a) the psychometric quality of PISA creative thinking assessment units in terms of validity, reliability, and comparability across participating countries; (b) the ability to construct a Creative Thinking scale and, possibly, subscales; (c) the inclusion of all the creative thinking units and forms in Field Trial 2021. It also generated (d) insights for further enrichment of the coder training materials utilized in coder training for Field Trial 2021 and the Main Survey 2022 (ACT, 2020<sup>[4]</sup>).

In 2021 a further Field Trial (FT) was conducted with 44 countries to provide additional evidence of the validity and reliability of the creative thinking assessment. Among the total of 38 CT items, two items were machine-scored, and the remaining 36 items were human-scored items. For the human-scored items in the 2020 LFT and the 2021 FT, all coding processes were performed by each country's coders. The ACT team provided national coder training and supported the national coding teams through a standard PISA query service. Items were initially reviewed for appropriateness (e.g. on task and on topic). Items determined to be appropriate were then scored using a single-digit or double-digit rubric. Scoring of Generate Creative Ideas and Evaluate and Improve Ideas items was conducted using a double-digit scoring rubric which captured data on the primary focus of a student response in addition to reflecting its credit level. Students demonstrated creativity in these facets by utilizing unconventional foci or employing innovative approaches. Scoring of the Generate Diverse Ideas items was conducted using a single-digit scoring rubric. Students demonstrated creativity in this facet by generating multiple, different ideas (see Annex Table 4.A.3).

### ***2020 Limited Field Trial Coder Training***

The coding guide for creative thinking was developed by test developers and performance scoring experts at ACT for the Field Trial with the support of the OECD. Coder training procedures and materials were informed by the cognitive labs and validation studies and included examples of genuine student responses. The English master version of the Creative Thinking Coding Guide was released in draft form prior to the in-person PISA International Coder Training meeting in January 2020.

Test developers and performance scoring experts from ACT, with the support of the OECD, facilitated discussions at that meeting. The coding guide used in the limited field trial was finalized based on these discussions. The updated English version of the coding guide and the French source version were released to countries in February 2020, prior to the beginning of the limited field trial data collection period.

### ***2021 Field Trial Coder Training***

The 2021 field trial International Coder Training for creative thinking was held over five days, virtually, due to the pandemic, in February 2021. Performance scoring experts from ACT developed online coding training modules and facilitated an interactive coder training webinar, held with representatives from participating 2021 Field Trial countries prior to coding. The training objectives included developing a foundational understanding of the construct and an in-depth understanding of the coding processes so that attending representatives would be prepared to train coders in their countries using the provided materials. In order to facilitate coder training, ACT's team developed comprehensive exemplar sets consisting primarily of selected authentic student responses intended to demonstrate a typical response for each

credit level and theme assignment (i.e. codes 00, 11, 12, 13, 21, 22, 23, and so on, with code 29 used to designate an unlisted theme, as explained in Section X). Discussion was also dedicated to reaching understanding and consensus about the coding rules for each item to better ensure consistency of coding within and between countries. Facilitators reviewed the layout of the coding guide, general coding principles, common problems, and guidelines for applying special codes. Workshop materials were optimized based on feedback from the LFT coder training, LFT coder queries, and translation referee updates to the 2021 coding guide. Workshop materials comprised primarily sample student responses that were provided for each item, and attendees were required to code them during the interactive workshop. Where there were disagreements about coding for an item, those were discussed in detail so that all attendees understood, and would be able to follow, the intent of the coding guides. In some instances, disagreements, particularly those highlighting possible cultural bias, led to modifications of the coding guide and/or workshop materials.

### **Preparation of data collection instruments**

#### *Preparing the Field Trial national student delivery systems (SDS)*

The process for creating the field trial national student delivery system (SDS) followed the approach used during the field trial, beginning with assembly and testing of the master SDS followed by the process for assembling national versions of the field trial SDS. After all components of national materials were locked, including the questionnaires and cognitive instruments, the student delivery system was assembled and tested first by Core A. 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 final systems were released for the Field Trial. PISA 2022 Creative Thinking was only administered on computer-based participants.

### **Field Trial Coding Procedures**

The FT design required that two independent coders review and code each student's responses at a credit level of either 0,1 (no credit or credit) or 0, 1, or 2, (no credit, partial credit, or full credit) thus generating inter-rater reliability at the credit level. In addition, two selected English-fluent bilingual coders from each country reviewed and coded 30 pre-designated anchor responses to verify coder reliability across the countries. These anchor responses were selected from earlier pilot studies conducted in Australia, Canada, Colombia, Singapore, and South Africa, and represented a range of responses at all credit levels (ACT, 2019<sup>[3]</sup>).

For the items measuring either the Generate Creative Ideas or the Evaluate and Improve Ideas facets, coders were required to use a second digit to indicate the primary theme of each response that earned partial or full credit.

Responses that received partial credit could only use values of 1-3 as the second digit to represent the preliminary conventional themes chosen based on available student responses (11, 12, or 13); however, responses that received full credit could use up to 9 different values for the second digit, with the ninth value representing all themes not associated with themes 1-8 (i.e. 21 through 29). The resulting data informed distinctions between "conventionality" and "unconventionality" of themes across a diverse international student cohort.<sup>1</sup>

Inter-rater reliability (IRR) on anchor responses across all items and coder pairs was 0.71. the average quadratic Kappa was also relatively high (0.79). Items were reviewed for the item category response functions, item quality. Items that exhibited high omit and not-reached rates were reviewed to rule out technical issues with the platform. Cluster placement was also considered to be a contributing factor when exploring reasons for high rates of omission or not reached coding. Items were further analysed for item difficulty, item discrimination, response time, position effect, IRT scaling, Item model fit, IRT parameters

and student theta estimates, evaluation of subscores on domain and facet levels, and differential item functioning via the item-total score curves from different country-by-language groups. The findings supported (a) the psychometric quality of PISA Creative Thinking assessment units in terms of validity, reliability, and comparability across participating countries; (b) the ability to construct a Creative Thinking scale; and (c) the inclusion of 20 of the 21 the Creative Thinking units in the 2022 Main Survey. For details of the findings please refer to the PISA 2022 Creative Thinking Field Trial Research Report (ACT, 2021<sup>[5]</sup>).

### ***Field Trial Coder Queries***

As was the case during previous cycles, Core A set up and maintained a coder query service for the 2020 and 2021 field trials. 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. Core B3 test developers and performance scoring experts from ACT reviewed and responded to queries specific to the Creative Thinking test developers.

In addition to responses to new queries, Core B3 curated a selection of queries to include in the Coder Query Log containing accumulated responses from previous cycles of PISA. This helped foster consistent coding of creative thinking items. The query log was regularly updated and posted for National Centres on the PISA Portal as new queries were received and processed.

### ***National item review following the Field Trial***

The item feedback process began in August 2021 and concluded in October 2021 and was conducted in two phases. Phase 1 occurred before countries received their Field Trial data and the Phase 2 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 turnaround 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 B3. 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 B3 and, where approved, implemented.

### ***Field Trial Outcomes***

The 2021 Field Trial data analyses addressed the issue of construct and score validity and reliability, within and across countries, in addition to differential item functioning. Items were analysed for Inter-rater reliability on anchor responses, inter-rater reliability on all responses, average Quadratic Kappa, item category response functions, item quality, item omit and not-reached rates, item difficulty, item discriminations, Item response time, position effect, IRT scaling, item model fit, IRT parameters and student theta estimates, evaluation of sub-scores on domain and facet levels and differential item functioning (DIF).

Flagged items were further reviewed in terms of their sample size, contents, translations, and coding guides (verified translation vs non-verified translation of coding guides), student responses (indications of misunderstanding), performance in alternative languages for that country, performance on similar items in assessment for that country/language, performance on the other items in that unit, additional item flags for that item, LFT data vs FT data, planned optimizations for that item (e.g. theme changes, coding optimizations, cluster placement). Due to the operational timeline in PISA 2022, it was not possible to include new items in the test after this phase, and no substantial modifications were made to existing test items. Poorly performing items were removed from the test item pool provided coverage of the domain was

not affected significantly. For the Creative Thinking test, one unit consisting of two items was removed. The PISA 2021 Field Trial also generated insights for further enrichment of the coder training materials, including the coding guide, towards the 2022 Main Survey. Substantial work was undertaken, including reviewing large amounts of student responses, additional frequency analysis of themes, and identification of instructions that caused coding issues by being absent, too vague, or too restrictive. This resulted in substantial modifications of the coding guide, including updates to conventional and unconventional themes, refinement of theme descriptions, increased representation of exemplar responses, and edits to item-specific instructions to facilitate effective and consistent coding (see Annex Table 4.A.4).

## Main Survey

The PISA 2022 Main Survey was conducted between March and December 2022. The majority of countries completed the Main Survey data collection by August. 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 B3 contractors worked with countries to resolve any remaining issues and prepare the national instruments for the main survey.

### *Item selection*

The PISA 2022 Field Trial provided evidence in support of the psychometric quality of PISA Creative Thinking assessment units in terms of validity, reliability, and comparability across participating countries. Improvements in performance for the 20 units included in the Main Survey are anticipated based on optimizations to the coding guide, coder trainings, and cluster arrangements. Maintaining the same range of contexts from the field trial to the main survey provided good continuity and kept a consistent representation of skills and domains. Clusters were created following the final item selection and balanced based on the coverage of cognitive processes, the discrimination and difficulty of the items, and the total number of units and items. The duration of each unit was between 5 and 15 minutes. The units were organized into five mutually exclusive 30-minute blocks or clusters. The clusters were rotated according to the integrated design presented in Chapter 2 of this Technical Report. The assessment aimed to achieve a good balance between units that situate creative thinking within the two thematic content areas and the four domains.

### *Review by the Creative Thinking Expert Group*

The Creative Thinking Expert Group reviewed the pilot study data, the approach to item selection, the content and balance of the clusters, and signed off on the selection.

## Preparation of data collection instruments

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

The process for creating the main survey national student delivery system (SDS) followed the approach 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 first by Core A. 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 final systems were released for the main study. PISA 2022 Creative Thinking was only administered on computers.



## Main survey coding

### *Main Survey Coder Training*

The Main Study International Coder Training for Creative Thinking was held in February 2022. Analysis of Field Trial responses and coder queries helped Performance scoring experts from ACT improve upon online coding training modules and other coder training materials. Additional sample responses were included in the coding guide to better illustrate different types of responses. Workshop materials were also enhanced to include additional authentic student responses that better illustrate the boundaries between full credit, partial credit (where appropriate) and no credit.

The process used for the Main Survey International Coder Training was similar to the 2021 Field Trial International Coder Training in that self-guided modules were completed before full-group discussions. The training objectives again included developing a foundational understanding of the construct and an in-depth understanding of the coding processes so that attending representatives would be prepared to train coders in their countries using the provided materials. Facilitators again reviewed the layout of the coding guide, general coding principles, common problems, and guidelines for applying special codes, and workshop materials for each item. Following the international coder training, additional revisions were made to the Creative Thinking Coding Guide in response to discussions that took place at the meeting.

### *Main Survey Coder Queries*

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 students' responses. Queries were reviewed, and responses were provided by domain-specific teams including test developers and coding experts. Core B3 test developers and performance scoring experts from ACT reviewed and responded to queries specific to the Creative Thinking test. Relevant queries were included in the Coder Query Log, a resource maintained by Core A and accessible by all participant NPMs in the PISA Portal.

## References

- ACT (2021), *PISA 2022 Creative Thinking Field Trial Research Report*, ACT, Iowa City, IA. [5]
- ACT (2020), *PISA 2022 Creative Thinking Limited Field Trial Research Report*, ACT, Iowa City, IA. [4]
- ACT (2019), *PISA 2021 Creative Thinking Validation Study Research Report*, ACT, Iowa City, IA. [3]
- ACT (2018), *PISA 2021 Creative Thinking Cognitive Lab Research Report*, ACT, Iowa City, IA. [2]
- OECD (2019), *PISA 2021 Creative Thinking Framework (Third Draft)*, OECD, Paris, <https://www.oecd.org/pisa/publications/PISA-2021-creative-thinking-framework.pdf>. [1]

## Notes

---

1. The conventionality or unconventionality of responses was determined by the originality of the response amongst those in the entire pool of responses (OECD, 2019<sup>[1]</sup>).

## Annex 4.A. Creative thinking items

**Annex Table 4.A.1. Chapter 4: Creative thinking assessment trials and main study**

Tables	Title
Table 4.A.2	Distribution of items by Facet and Domain
Table 4.A.3	Creative Thinking Assessment Field Trial item distribution by facet, unit, and domain
Table 4.A.4	Creative Thinking Assessment Main Study item distribution by facet, unit, and domain

**Annex Table 4.A.2. Distribution of items by Facet and Domain**

Domain	Facet		
	Generate Diverse Ideas	Generate Creative Ideas	Evaluate & Improve Ideas
Visual Expression	2	2	4
Written Expressions	4	6	2
Social Problem Solving	4	3	3
Science Problem Solving	4	1	3

**Annex Table 4.A.3. Creative Thinking Assessment Field Trial item distribution by facet, unit, and domain**

Domain	Unit	Facet		
		Generate Diverse Ideas	Generate Creative Ideas	Evaluate and Improve Ideas
Visual	Unit 1		X	X
	Unit 2	X		X
	Unit 3		X	X
	Unit 4	X		X
Written	Unit 5	X	X	
	Unit 6	X	X	
	Unit 7	X	X	X
	Unit 8	X	X	X
	Unit 9		X	
	Unit 10		X	
Social	Unit 11	X	X	X
	Unit 12	X	X	
	Unit 13	X		X
	Unit 14	X		
	Unit 15			X
	Unit 16		X	
Science	Unit 17	X		
	Unit 18			X

	Unit 19	X	X	
	Unit 20	X		X
	Unit 21	X		X

**Annex Table 4.A.4. Creative Thinking Assessment Main Study item distribution by facet, unit, and domain**

Domain	Unit	Facet		
		Generate Diverse Ideas	Generate Creative Ideas	Evaluate and Improve Ideas
Visual	Unit 1		X	X
	Unit 2	X		X
	Unit 4	X		X
Written	Unit 5	X	X	
	Unit 6	X	X	
	Unit 7	X	X	X
	Unit 8	X	X	X
	Unit 9		X	
	Unit 10		X	
Social	Unit 11	X	X	X
	Unit 12	X	X	
	Unit 13	X		X
	Unit 14	X		
	Unit 15			X
	Unit 16		X	
Science	Unit 17	X		
	Unit 18			X
	Unit 19	X	X	
	Unit 20	X		X
	Unit 21	X		X

# 5 Context Questionnaire Development

## Introduction

This chapter describes the PISA 2022 context questionnaire development process, as guided by the 2022 framework, as well as its linking to questionnaires from previous PISA cycles of the PISA assessment, as set out in the PISA 2012, 2015, and 2018 questionnaire frameworks (OECD, 2013<sup>[1]</sup>; 2017<sup>[2]</sup>; 2019<sup>[3]</sup>). The constructs that need to be covered for monitoring trends in education are discussed in the context of research into the effectiveness of education systems. These measures have been used previously in PISA reports, as international indicators published in *Education at a Glance*, and in secondary analyses. For more information about the PISA Questionnaire Development, see OECD (2023<sup>[4]</sup>).

One of the major features of the implementation of PISA is the cyclical change in focus of the cognitive assessment: mathematics was the major domain of assessment in PISA 2003 and 2012 and is so again in PISA 2022, whilst reading literacy was the major domain of PISA 2000, 2009 and 2018, and science in PISA 2006 and 2015. The major domain of the cognitive assessment is also the focus of domain-specific context assessment in the associated questionnaire – in other words, various mathematics-related constructs were assessed in the PISA 2022 questionnaire since mathematics was the major domain. However, there is also a need for stability in measures administered in different cycles in order to gauge and understand trends in education. Stability has to be considered at two levels: across periods of three years (various questions in the questionnaires tend to recur in every cycle) and in subject-specific constructs across periods of nine years (mathematics-specific constructs assessed in the 2012 wave could be reused in 2022)<sup>1</sup>.

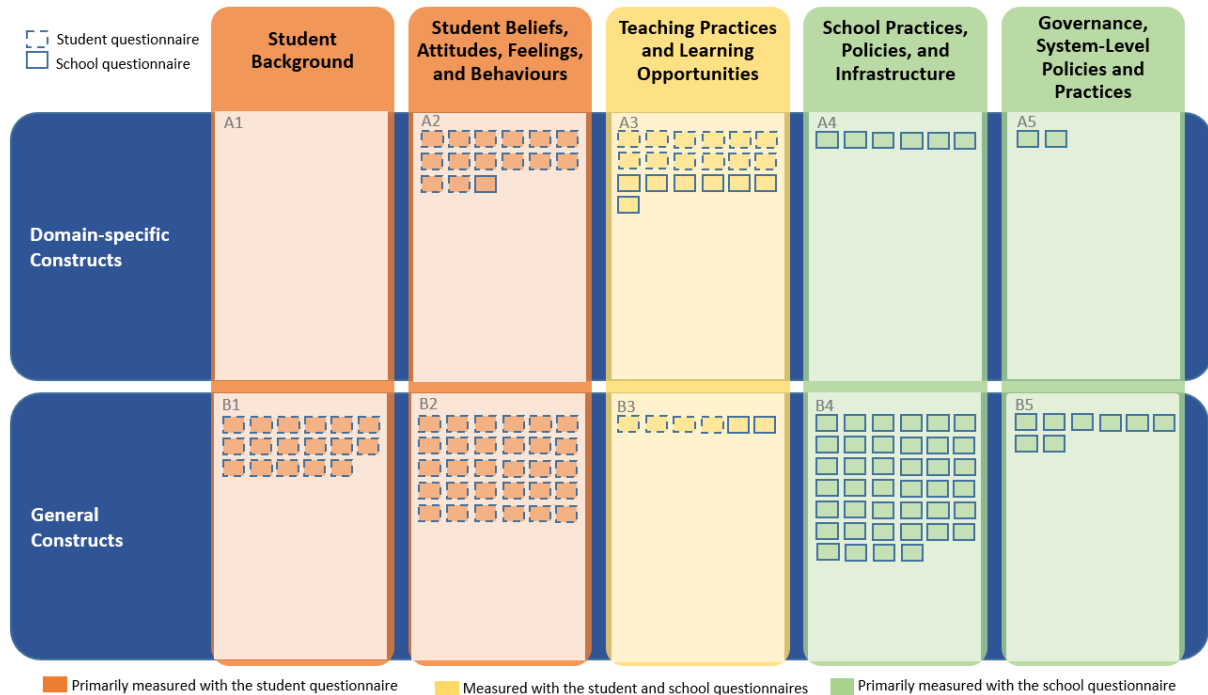
## The role of the PISA context questionnaire framework in development

The PISA 2022 two-dimensional framework taxonomy is presented in Figure 5.1. The first dimension classifies proposed constructs into the two overarching categories distinguished by the PISA Governing Board (PGB; domain-specific constructs and general constructs, with the latter including Economic, Social, and Cultural Status [ESCS]). The second dimension classifies proposed constructs into five categories based on key areas of educational policy setting at different levels of aggregation (Student Background; Student Beliefs, Attitudes, Feelings, and Behaviours; Teaching Practices and Learning Opportunities; School Practices, Policies, and Infrastructure; and Governance, System-Level Policies and Practices). The small boxes in the taxonomy below indicate the relative distribution of constructs in the PISA 2022 context questionnaires across all modules described in this framework.

Every module represents a focus around a topic, and the set of 21 content modules (see Annex Table 5.A.2) covers a wide and comprehensive array of educational policy issues that are relevant across all participating countries/economies. The framework first discusses student background constructs, followed by student beliefs, attitudes, feelings, and behaviours constructs, teaching and learning constructs, and finally school policy and governance constructs. PISA treats the mandatory core

questionnaires (school questionnaire and student questionnaire) separately from the optional questionnaires, which countries must opt into.

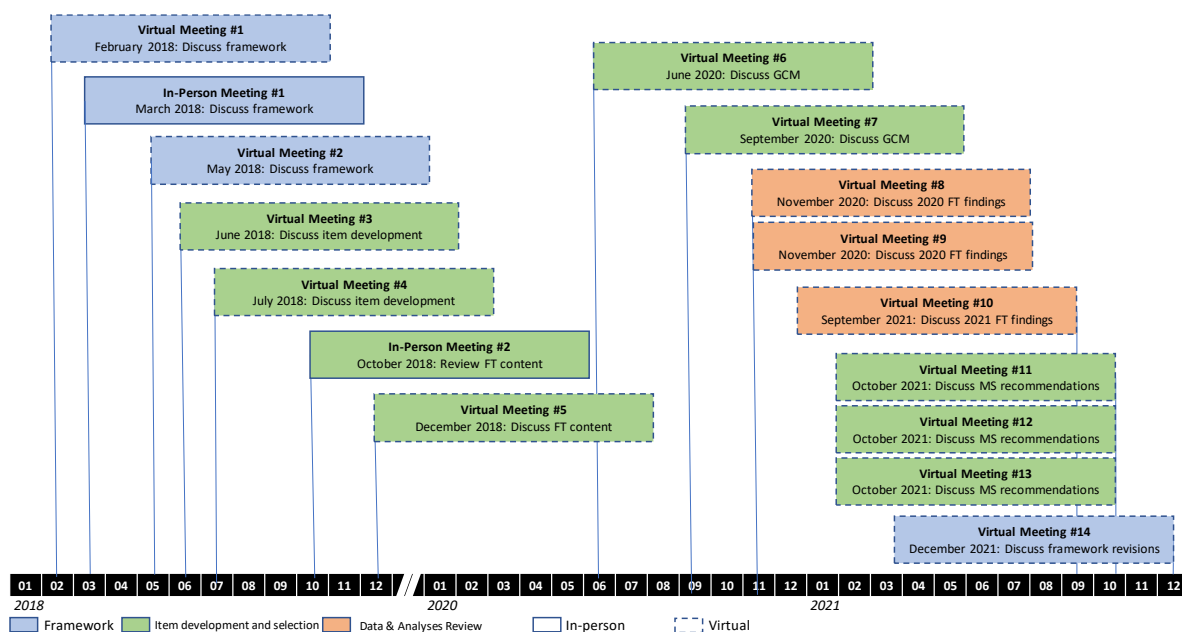
**Figure 5.1. PISA 2022 Questionnaire Framework and Modules**



As reflected in Figure 5.1, the PISA 2022 questionnaires have a stronger focus on general constructs (including economic, social, and cultural status) compared to domain-specific constructs. This was in response to the PGB's recommendation to re-balance questionnaire content in the direction of a larger focus on general constructs and a reduced focus on domain-specific constructs.

As in previous cycles, the Questionnaire Expert Group (QEG) guided the development of the PISA context questionnaires and framework through regular meetings. QEG members reviewed drafts of each instrument as well as feedback from countries and economies and discussed the material together with the OECD Secretariat and the international contractors to ensure the concordance between the assessment, the context questionnaires, and the corresponding frameworks. During this process, the QEG for PISA 2022 liaised with the Mathematics Expert Group (MEG) and received and reacted to presentations from the Creative Thinking contractor, guaranteeing a close link between the development of the assessment framework and tests and the questionnaire development process. Figure 5.2 provides an overview of the junctions at which the QEG was consulted via in-person or virtual meetings. Please note, meetings 11 through 13 were originally planned as a single in-person meeting but facilitated as a series of shorter virtual meetings due to COVID-19 travel restrictions.

Figure 5.2. Virtual and in-person meetings with the PISA 2022 Questionnaire Expert Group (QEG)



## Questionnaires for different respondent groups

There were seven context questionnaires administered in PISA 2022. Two of them, the student and school questionnaires, were considered core questionnaires and were administered in all participating countries/economies. The other five questionnaires were optional and administered in a subset of the participating countries/economies to students, their parents or guardians, and teachers. Optional questionnaires for student respondents were administered in the order as described below, immediately after the STQ.Core Context Questionnaires

**Student Questionnaire (StQ).** The 35-minute PISA Student Questionnaire was administered to all students participating in the PISA assessment. A complete version was administered to those taking the assessment on computer, while countries/economies testing on paper administered a paper version containing a subset of the questions. The computer-based version of the StQ further utilized a new within-construct matrix sampling questionnaire design, where each student received a random selection of five questions about the same topics or “constructs” from a “pool” of approximately ten questions for most constructs. This design, which was developed based on a series of methodological studies (Bertling and Weeks, 2018<sup>[5]</sup>; 2020<sup>[6]</sup>) with guidance from the PISA Technical Advisory Group, maximizes the number of policy-relevant questions that can be used in the student questionnaire without increasing individual student response burden. Annex Table 5.B.1 of this chapter lists the questions included in the student background questionnaire, the module and construct they measure, and whether they were administered as part of the PBA assessment or using matrix-sampling.

Details regarding the creation of scaled indices based on this new design can be found in the Chapter 18 of this report.

**School Questionnaire (ScQ).** The 45-minute PISA School Questionnaire was administered to the principals of the schools with students participating in PISA. It was administered on computer in countries taking the assessment on computer, while countries/economies using paper-based testing administered a paper version of the same questionnaire.

### ***Optional Context Questionnaires***

**Financial Literacy Questionnaire (FLQ).** This 10-minute computer-based questionnaire was administered to all participating students in countries/economies that were taking the assessment on computer and administered the Financial Literacy assessment. It included questions about students' access to financial information and education as well as their practical financial experiences.

**Information Communication Technology Questionnaire (ICQ).** This 10-minute computer-based questionnaire was administered to all participating students in countries/economies that were taking the assessment on computer and chose to implement this option. It included questions about students' usage of electronic and digital devices, as well as their confidence and attitudes towards ICT.

**Well-being Questionnaire (WBQ).** This 10-minute computer-based questionnaire was administered to all participating students in countries/economies that were taking the assessment on computer and chose to implement this option. It included questions about students' health and well-being, as well as activities with friends and family.

**Parent Questionnaire (PaQ).** This 30-minute paper-based questionnaire was administered to parents or guardians of all participating students in countries/economies that chose to implement this option. It included questions about learning contexts, support, and resources at home as well as spending on education and parents' or guardians' mathematics-related interests and attitudes.

**Teacher Questionnaire (TQ).** This 40-minute computer-based questionnaire was administered to teachers in countries/economies that chose to implement this option. It was administered as an integrated questionnaire that utilized digital routing to direct respondents to either a mathematics teacher or a general teacher module. After completing the initial module, all respondents then received a creative thinking module and a teacher well-being module.

Annex Table 5.A.3 provides an overview how each of these seven questionnaires relates to the educational policy areas outlined in the framework.

## **Phases of Questionnaire Development and QUALITY ASSURANCE**

Questionnaire development for PISA 2022 followed a multi-step process including several defined interaction points with subject matter experts, respondent groups, and stakeholders, and defined mechanisms to ensure quality of the developed instruments and comparability of the data across countries/economies. The following sections each give a short summary of each questionnaire development phase alongside relevant quality assurance strategies associated with each phase.

### ***Development of initial item pool***

Questionnaire development started with evaluating the existing questionnaire pool for PISA and identifying areas that required new development based on the PISA 2022 context questionnaire framework. Following prioritization with the QEG and the OECD, new questions for all questionnaires except for the WBQ, which was administered without changes from the PISA 2018 version, were drafted based on principles outlined in the framework.

### ***Small-scale pre-testing in cognitive interviews***

A subset of all newly-developed questionnaire material for the StQ representing a range of cognitive and language complexity was pre-tested in small samples of students in Hong Kong, China, India, and Brazil<sup>2</sup> during the development stage. The small-scale pre-testing was conducted in Cantonese, Hindi, and Portuguese in an effort to widen the languages included in pre-testing beyond western languages. Pre-



testing took the form of two rounds of in-person one-on-one cognitive interviews and a third round of virtual one-on-one interviews for the Global Crises Module (see below), each with small groups of students from socioeconomically diverse backgrounds. Interviews were facilitated under general leadership of the PISA Core A contractor by teams led by members of the QEG, to collect feedback from respondents representing diverse geographic, linguistic, and cultural backgrounds. During each cognitive interview session, an interviewer provided students, in paper-based format, with a set of thematically-grouped questions. Students were asked to provide answers to all questions in the set. When the student was finished providing their answers, the interviewer asked a series of retrospective probes associated with each question in the set. These probes asked about students' interpretation of the question; their understanding of words in specific items of a matrix question; other words or parts of the question that they found confusing; and the overall level of difficulty they reported in answering the question. Once the student finished responding to the probes, the interviewer provided the student with another set of questions to answer.

In the first round of cognitive interviews, four thematically-defined sets of questions were tested among student respondents. In the second round, another five thematically-defined sets were tested. A second goal of the cognitive interviews was to collect data on students' understanding of different response options (i.e. agreement, like-me, and frequency type response options) to guide recommendations regarding which response options to use for specific questionnaire content in PISA 2022. Two additional types of activities were performed during the cognitive interviews as preliminary steps toward response option classification for PISA 2022: card-sorting exercises, and response option comparisons.

### ***Feedback from participating countries/economies***

All newly-developed material was shared with representatives of countries/economies at an early stage in the development process to obtain in-depth feedback. National Centres were asked for ratings on several important factors for each question to be implemented in PISA, including 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. students, teachers, principals) is indeed the best respondent group to answer the question. In this context, a very important aspect of ratings touched on issues of sensitivity. Feedback was collected on whether a topic might be sensitive, 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/economies were asked to give an overall rating of each proposed question and provide any additional comments or concerns that might improve the material. A similar review was repeated after the international Field Trial (FT).

### ***Translatability assessment***

To enhance comparability, a translatability assessment of the questionnaire material was carried out before finalizing instruments for the FT. Linguistic experts evaluated the material with due consideration for the Ask-the-Same-Question (ASQ) model (Harkness, 2003<sup>[7]</sup>). This approach seeks to optimize the wording in the source questionnaire so that the items can be translated into 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 (Cantonese), Slavic languages (Bosnian, Croatian, Russian), an Indo-German language (German), a Romance language (French, Portuguese), Turkic (Turkish), and Finno-Ugric (Hungarian). 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.

### ***Refinement of item pool and creation of international master version for FT***

After cognitive interviews, feedback from the review by countries/economies, conclusion of the translatability assessment, and review by the PISA subject matter expert groups (i.e. QEG, MEG, Creative Thinking Expert Group - CTEG), the item pool was revised for administration in the FT. An important addition to the questionnaires at this point was the Global Crises Module (GCM) (Bertling et al., 2020<sup>[8]</sup>). The GCM was developed as an additional questionnaire module for student and school questionnaire respondents with a focus on effects of the COVID-19 pandemic on student learning and well-being and the degree of interruptions or changes to education across participating countries/economies. Please note, although the GCM was added to the development process at a later stage than other questionnaire materials, the questions went through the same quality assurance steps as all other materials.

### ***Centralised trend material transfer from previous PISA cycles***

For the computer-based questionnaires, in earlier PISA cycles 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 2022 was centrally transferred within the electronic platform. Because the process for adapting and translating questionnaires this cycle required that all adaptations were documented in English in the electronic platform before being translated, when the contractors transferred trend material they also supplied the English back-translation of the trend text, which the country/economy confirmed during their review. Any changes to these trend questions needed to be requested and justified by the country/economy. This process allowed for external control to preserve national trend material from the previous cycle in PISA 2022.

For the paper-based questionnaires, the international contractors did not perform a centralized transfer of trend material. Participating countries/economies were provided with their questionnaires from the previous cycle of PISA (if they participated) and were asked to copy the trend items into the PISA 2022 questionnaires.

### ***Adaptation negotiation and verification of all questionnaire material***

In some cases, cultural traditions, local understanding of a question or features of the education system vary largely, leading to the need for adaptations to the questionnaires. As in previous PISA cycles, the National Centres in each country/economy were asked to document which 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 negotiation of adaptations and translation of the customized national text into the local language, 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 National Centre, and the linguistic quality control team. The chapter on translation verification in this Technical Report has additional information about this process. All final questionnaire material was then implemented into the paper-based or computer-based versions, tested in the system, and provided to the PISA participants.

### ***Large-scale testing in international Field Trial***

All question developed for potential inclusion in the PISA 2022 MS, including the GCM, were administered to the respective respondent group in the PISA 2022 international FT. In addition to examining each question's performance across participating countries/economies, several methodological experiments were conducted as part of the FT, in an interest of choosing the most appropriate operationalisation for each construct described in the PISA 2022 Questionnaires Framework. These experiments comprised

comparison of multiple choice (MC) and fill-in questions, comparison of agreement and frequency response options, comparison of abstract and concrete frequency response options, and comparison of mother/father-focused with parent or guardian-focused education- and occupation-related questions. Results for each experiment were discussed with relevant PISA expert groups and the OECD secretariat prior to determining the final direction with questionnaire selection for the Main Survey (MS).

### ***Finalization of item pool for international Main Survey***

A reduction of questions was needed across all questionnaires from the FT to the MS, except for the WBQ, which was administered without changes from the PISA 2018 version. Item recommendations and subsequent decisions for the MS instruments were based on the empirical performance of the items based on data from the first batch of countries/economies with submitted FT data as well as a consideration of redundancies and framework coverage and consultation with key stakeholders, including the QEG, MEG, CTEG, as well as National Centres in each country/economy. Based on findings from the above-mentioned methodological experiments, it was determined that the PISA 2022 MS would retain the mother-father focused fill-in question format from previous cycles for occupation-related questions, that agreement-types response options would be used for General Social and Emotional Characteristics, thereby maximising consistency with the OECD's survey on social and emotional skills (SSES), and that newly-developed frequency questions would use more concrete instead of highly abstract response options in efforts to improve cross-country comparability.

### ***Main Survey review by countries***

Between the FT and MS each National Centre was asked to review its FT data for unexpected response distributions to the questions and to investigate whether the data indicated that there were any errors in the adaptations they requested or the translations of the questionnaires that needed to be corrected. This included updates due to errata. All requested changes were checked by the international contractors and documented in the QAS. Approved changes to translation were implemented by verifiers.

All final questionnaire material was then implemented into the paper-based or computer-based versions, tested, and provided to the PISA participants in advance of the MS. More details about the preparation of the questionnaires is included in Chapter 19.

## **Summary**

Each of the steps in this development process ensured that questions included in PISA 2022 were systematically evaluated and iteratively refined based on insights from empirical data before the finalisation of the international versions of the questionnaires. See Chapter 19 for how the questionnaire design was implemented in the system and see Chapter 18 for how derived variables for reporting were created for the questionnaires.

## References

- Bertling, J. et al. (2020), “A tool to capture learning experiences during COVID-19: The PISA Global Crises Questionnaire Module”, *OECD Education Working Papers*, No. 232, OECD Publishing, Paris, <https://doi.org/10.1787/9988df4e-en>. [8]
- Bertling, J. and J. Weeks (2020), *Getting More Bang for Your Buck: Within-construct Questionnaire Matrix Sampling*, Paper presented to PISA Technical Advisory Group, September 2020, Princeton, NJ. [6]
- Bertling, J. and J. Weeks (2018), *Plans for Within-construct Questionnaire Matrix Sampling in PISA 2021*, Paper presented to PISA Technical Advisory Group, August 2018, Princeton, NJ. [5]
- Harkness, J. (2003), “Questionnaire Translation”, in Harkness, J. (ed.), *Cross-Cultural Survey Methods*, Wiley, Hoboken. [7]
- OECD (2023), *PISA 2022 Context Questionnaire Framework: Balancing Trends and Innovation*, OECD, Paris, [https://www.oecd-ilibrary.org/fr/education/pisa-2022-assessment-and-analytical-framework\\_dfe0bf9c-en](https://www.oecd-ilibrary.org/fr/education/pisa-2022-assessment-and-analytical-framework_dfe0bf9c-en). [4]
- OECD (2019), “PISA 2018 Questionnaire Framework”, in *PISA 2018 Assessment and Analytical Framework*, OECD Publishing, Paris, <https://doi.org/10.1787/850d0ef8-en>. [3]
- OECD (2017), “PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving”, in *PISA 2015 Context Questionnaires Framework*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264281820-7-en>. [2]
- OECD (2013), “Context Questionnaires Framework”, in *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264190511-en>. [1]

## Notes

- 
1. There is a four-year gap between the last and the current PISA cycles (i.e. 2018 and 2022) and a ten-year gap (2012 and 2022) between the last two cycles focused on mathematics due to a one-year delay as a result of the COVID-19 pandemic.
  2. We thank Wilima Wadhwa, Kit-Tai Hau, and Ricardo Primi and their teams for their dedication and support in facilitating these studies.

# Annex 5.A. PISA 2022 Questionnaire Framework Content Modules

**Annex Table 5.A.1. Chapter 5: Content Modules and Policy Categories**

Tables	Title
Table 5.A.2	Content Modules defined in PISA 2022 Questionnaire Framework
Table 5.A.3	Overview of the five categories based on key areas of educational policy setting at different levels of aggregation in the PISA 2022 framework covered by the questionnaires

**Annex Table 5.A.2. Content Modules defined in PISA 2022 Questionnaire Framework**

No.	Module	No.	Module
1	Basic Demographics	11	School Type and Infrastructure
2	Economic, Social, and Cultural Status (ESCS)	12	Selection and Enrolment
3	Educational Pathways and Post-Secondary Aspirations	13	School Autonomy
4	Migration and Language Exposure	14	Organisation of Student Learning at School
5	PISA Preparation and Effort	15	Exposure to Mathematics Content
6	School Culture and Climate	16	Mathematics Teacher Behaviours
7	Subject-specific Beliefs, Attitudes, Feelings, and Behaviours	17	Teacher Qualification, Training, and Professional Development
8	General Social and Emotional Characteristics	18	Assessment, Evaluation and Accountability
9	Health and Well-being	19	Parental/Guardian Involvement and Support
10	Out-of-school Experiences	20	Creative Thinking
		21	Global Crises

**Annex Table 5.A.3. Overview of the five categories based on key areas of educational policy setting at different levels of aggregation in the PISA 2022 framework covered by the questionnaires**

	Main Survey Length (minutes)	Framework Coverage				
		Student Background	Student Beliefs, Attitudes, Feelings, and Behaviours	Teaching Practices and Learning Opportunities	School Practices, Policies, and Infrastructure	Governance, System-Level Policies and Practices
Student	35	√	√	√		
School	45		√	√	√	√
Financial Literacy	10	√	√	√		
ICT	10	√	√	√	√	
Well-being	10	√	√	√		
Parent	30	√		√	√	
Teacher	40			√	√	

## Annex 5.B. Student Questionnaire

Annex Table 5.B.1. Details of PISA 2022 Student Questionnaire Main Survey questions

Question No.	Module	Construct	Within-construct matrix sampling (CBA only)	In PBA
ST001	Basic demographics	Grade	no	yes
ST003	Basic demographics	Date of birth	no	yes
ST004	Basic demographics	Gender	no	yes
ST002	Educational career	Current study programme	no	yes
ST250	Economic, social, and cultural status (ESCS)	Home possessions	no	yes
ST251	Economic, social, and cultural status (ESCS)	Home possessions	no	yes
ST253	Economic, social, and cultural status (ESCS)	Digital devices in the home	no	yes
ST254	Economic, social, and cultural status (ESCS)	Digital devices in the home	no	yes
ST255	Economic, social, and cultural status (ESCS)	Books in the home	no	yes
ST256	Economic, social, and cultural status (ESCS)	Books in the home	no	no
ST230	Basic demographics	Number of siblings	no	yes
ST005	Economic, social, and cultural status (ESCS)	Mother's education	no	yes
ST006	Economic, social, and cultural status (ESCS)	Mother's education	no	yes
ST007	Economic, social, and cultural status (ESCS)	Father's education	no	yes
ST008	Economic, social, and cultural status (ESCS)	Father's education	no	yes
ST014	Economic, social, and cultural status (ESCS)	Mother's occupation	no	yes
ST015	Economic, social, and cultural status (ESCS)	Father's occupation	no	yes
ST258	Economic, social, and cultural status (ESCS)	Food insecurity	no	yes
ST259	Economic, social, and cultural status (ESCS)	Subjective socioeconomic status	no	yes
ST019	Migration and language exposure	Immigration background	no	yes
ST021	Migration and language exposure	Immigration background	no	yes
ST022	Migration and language exposure	Primary home language	no	yes
ST226	Educational career	Time attended current school	no	yes
ST125	Educational career	Age started ISCED 0	no	yes
ST126	Educational career	Age started ISCED 1	no	yes
ST127	Educational career	Grade repetition	no	yes
ST260	Educational career	Truancy	no	yes
ST261	Educational career	Truancy	no	yes
ST062	Educational career	Truancy	no	yes
ST267	School culture and climate	Quality of student-teacher relationships	yes	yes
ST034	School culture and climate	Sense of belonging	yes	yes

Question No.	Module	Construct	Within-construct matrix sampling (CBA only)	In PBA
ST038	School culture and climate	Being bullied	no	yes
ST265	School culture and climate	Feeling safe	no	yes
ST266	School culture and climate	School safety risks	no	yes
ST294	Out-of-school experiences	Activities before school	no	yes
ST295	Out-of-school experiences	Activities after school	no	yes
ST326	Health and well-being	Time spent on online activities	no	yes
ST322	Health and well-being	Digital device usage behaviours	yes	no
ST307	General social and emotional characteristics	Perseverance	yes	no
ST309	General social and emotional characteristics	Self control	yes	yes
ST301	General social and emotional characteristics	Curiosity	yes	yes
ST343	General social and emotional characteristics	Cooperation	yes	no
ST311	General social and emotional characteristics	Empathy	yes	no
ST315	General social and emotional characteristics	Trust	yes	yes
ST303	General social and emotional characteristics	Perspective taking	yes	no
ST305	General social and emotional characteristics	Assertiveness	yes	yes
ST345	General social and emotional characteristics	Stress resistance	yes	no
ST313	General social and emotional characteristics	Emotional control	yes	yes
ST263	Subject-specific beliefs, attitudes, feelings, and behaviours	Growth mindset	no	no
ST016	Health and well-being	Overall life satisfaction	no	yes
ST059	Organization of student learning at school	Class periods per week in in mathematics	no	yes
ST296	Out-of-school experiences	Time spent on mathematics homework	no	yes
ST272	Mathematics teacher behaviours	Perceived quality of mathematics instruction	no	yes
ST273	Mathematics teacher behaviours	Disciplinary climate in mathematics	yes	yes
ST270	School culture and climate	Mathematics teacher support	no	yes
ST285	Mathematics teacher behaviours	Cognitive activation in mathematics: Foster reasoning	yes	yes
ST283	Mathematics teacher behaviours	Cognitive activation in mathematics: Encourage mathematical thinking	yes	no
ST275	Exposure to mathematics content	Exposure to formal and applied mathematics tasks	yes	yes
ST276	Exposure to mathematics content	Exposure to mathematics reasoning and 21st century mathematics topics	yes	yes
ST268	Subject-specific beliefs, attitudes, feelings, and behaviours	Preference of math over other core subjects, and Perception of mathematics as easier than other subjects	no	no
ST290	Subject-specific beliefs, attitudes, feelings, and behaviours	Mathematics self-efficacy: formal and applied mathematics	yes	yes
ST291	Subject-specific beliefs, attitudes, feelings, and behaviours	Mathematics self-efficacy: reasoning and 21st century mathematics	yes	no
ST289	Subject-specific beliefs, attitudes, feelings, and behaviours	Subjective familiarity with mathematics concepts	yes	no

Question No.	Module	Construct	Within-construct matrix sampling (CBA only)	In PBA
ST293	Subject-specific beliefs, attitudes, feelings, and behaviours	Proactive mathematics study behavior	yes	yes
ST292	Subject-specific beliefs, attitudes, feelings, and behaviours	Mathematics anxiety	yes	yes
ST297	Out-of-school experiences	Participation in additional mathematics instruction	no	yes
ST334	Creative thinking	Creative self-efficacy	yes	no
ST335	Creative thinking	Creative school and class environment	yes	no
ST336	Creative thinking	Creative peers and family environment	yes	no
ST337	Creative thinking	Creative school activities	no	no
ST338	Creative thinking	Creative outside school activities	no	no
ST339	Creative thinking	Beliefs about creativity	no	no
ST340	Creative thinking	Creativity and openness to intellect	yes	no
ST341	Creative thinking	Openness to art and reflection	no	no
ST342	Creative thinking	Imagination and adventurousness	yes	no
ST300	Parental/guardian involvement and support	Family support	yes	yes
ST327	Post-secondary preparedness and aspirations	Expected educational level	no	yes
ST329	Post-secondary preparedness and aspirations	Expected occupation	no	yes
ST330	Post-secondary preparedness and aspirations	Future study or work information	yes	no
ST324	Post-secondary preparedness and aspirations	Outlook on future educational career	yes	no
ST347	Global Crises	Type/duration of school closure	no	yes
ST348	Global Crises	School actions/activities to sustain learning	yes	yes
ST349	Global Crises	Type of digital device used for school work	no	yes
ST350	Global Crises	Subjective impression of learning during school closure	no	yes
ST351	Global Crises	Types of learning resources used while school was closed	yes	yes
ST352	Global Crises	Problems with self-directed learning	yes	yes
ST353	Global Crises	Family support for self-directed learning	yes	yes
ST354	Global Crises	Feelings about learning during school closure	yes	yes
ST355	Global Crises	Self-directed learning self-efficacy	yes	yes
ST356	Global Crises	Feeling of preparedness for future school closures	no	yes



# 6 Sample Design

## Target population and overview of the sampling design

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

- 15-year-old students enrolled full-time in educational institutions
- 15-year-old students 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/economy (as well as students from other countries/economies attending any of the programmes in the first three categories).

It was recognised that no testing of 15-year-old students schooled in the home, workplace or out of the country/economy would occur and therefore these 15-year-olds are 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 56-day period, referred to as the testing period, between March 1<sup>st</sup>, 2022 and October 31<sup>st</sup>, 2022, 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/economies. As most 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/economies testing in April 2022, the target population could have been defined as all students born in 2006 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/economy testing in March or in May to still define the national target population as all students born in 2006. If the testing took place between June and December, the birth date definition had to be adjusted so that in all countries/economies 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.

The situation with the COVID-19 pandemic made it difficult for several countries to adhere strictly to the testing period and the age definition for the target population just discussed. Recognizing the challenges of conducting assessments in such an environment, it was proposed by the international consortium that certain minor violations of these standards be sanctioned in advance, so that countries did not face uncertainty as they incurred the cost and burden of conducting the assessments. Thus, for PISA 2022, the

OECD and the PISA Technical Advisory Group accepted the following types of deviations from the standards:

- a. Extension of the assessment period beyond 56 days, where students remain within the PISA-eligible age range, would be agreed to with the OECD's implicit approval.
- b. Extension of the assessment period that would not exceed the allowed 56 days, but would result in some assessed students who are outside of the PISA-eligible age range **by less than a week**, would be agreed to with the OECD's implicit approval.
- c. Extension of the assessment period that would **both** exceed 56 days AND result in assessed students who are outside of the PISA eligible age range would require further consultation with the contractors and the OECD before approval of such a deviation would be granted.

In all countries/economies, the default 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 type of sampling is referred to as systematic with probability proportional to size (PPS) sampling. Prior to selecting them, schools in the sampling frame were assigned to mutually exclusive groups based on school characteristics called explicit strata. These were formed to improve the precision of sample-based estimates. Stratification variables for each country/economy are presented in Annex Table 6.A.2.

The second-stage sampling units in countries/economies 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. Countries/economies participating in the computer-based assessment (CBA) had to set a target cluster size (TCS) of 42 students, while countries/economies participating in the paper-based assessment (PBA) had to set a TCS of 35 students. Variations to the TCS were allowed in consultation with the sampling contractors for factors such as expected student nonresponse.

The sample size within schools is prescribed, within limits, in the PISA Technical Standards (see Annex I). From each list of eligible students within a school that contained more than the target cluster size, a sample of around 42 (or 35 for the case noted above) students were selected with equal probability, and for lists with fewer than the target number, all students on the list were selected.

The students selected for financial literacy were an additional sample of students above and beyond those needed for PISA. This was the same approach used in 2018.

## 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 probabilistic sampling in a way that ensured representation of the full target population of 15-year-old students in the participating countries/economies.

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

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

National Project Managers (NPMs) might have found it unavoidable to reduce their coverage of the target population by excluding, for instance, a small, remote geographical region due to inaccessibility, or language differences, possibly due to political, organisational or operational reasons, or presence of special education needs students. Areas deemed to be part of a country/economy that included students in the PISA target population, but which were not included for sampling, 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/economies 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 2022 Technical Standards (see Annex I).

Exclusion could also take place either at the school level (exclusion of entire schools) or at the within-school level (exclusion of individual students). These exclusions were often for special education needs or language differences.

International within-school exclusion of students was allowed for the following groups:

- Intellectually disabled students: these students who have a documented 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: these 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 experience in the language of assessment: these 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 taught in a language of instruction for the main domain for which no materials were available. PISA Technical 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/economy 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.
- 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 and the OECD. A specific subgroup of students (i.e., students with severe dyslexia, dysgraphia, or dyscalculia) could be identified for whom exclusion was necessary but for whom the first three within-school exclusion categories did not explicitly apply, so that a more specific within-school exclusion definition was needed.
- Students currently not attending in-person classes, receiving all their instruction online/virtually and not coming to schools for tests/assessments. This exclusion type was exceptionally added for PISA 2022 due to the coronavirus pandemic.

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/economy (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. 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 all be excluded in the field at the 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 student population. 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/economy'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 notate the results, or take other action in relation to reporting the data.

### ***Accuracy and precision***

A minimum of 150 schools was selected in each country/economy, but if a participating country/economy had fewer than 150 schools in existence, then all schools were selected for participation. Within each participating school, a predetermined number of students – the target cluster size, as defined earlier – was 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 6 300 assessed students was needed in computer-based countries/economies, or 5 250 assessed students in paper-based countries/economies. In cases where the entire population had fewer students, all students were selected. It was possible to negotiate a target cluster size that differed from 42 students (or 35 as noted above). When this was the case, the sample size of schools was increased to more than 150 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 25 students to ensure adequate accuracy in estimating variance components within and between schools – a major analytical objective of PISA.

Countries/economies doing the FL option needed an additional 1 650 assessed students for FL. To accomplish this, the target cluster size was usually increased for countries/economies participating in the financial literacy assessment. For example, a county/economy that would have sampled 42 students in each school generally increased its TCS to 53 to accommodate the financial literacy sample. In some instances, the country/economy opted to increase the school sample size to achieve the required number of students selected for financial literacy.

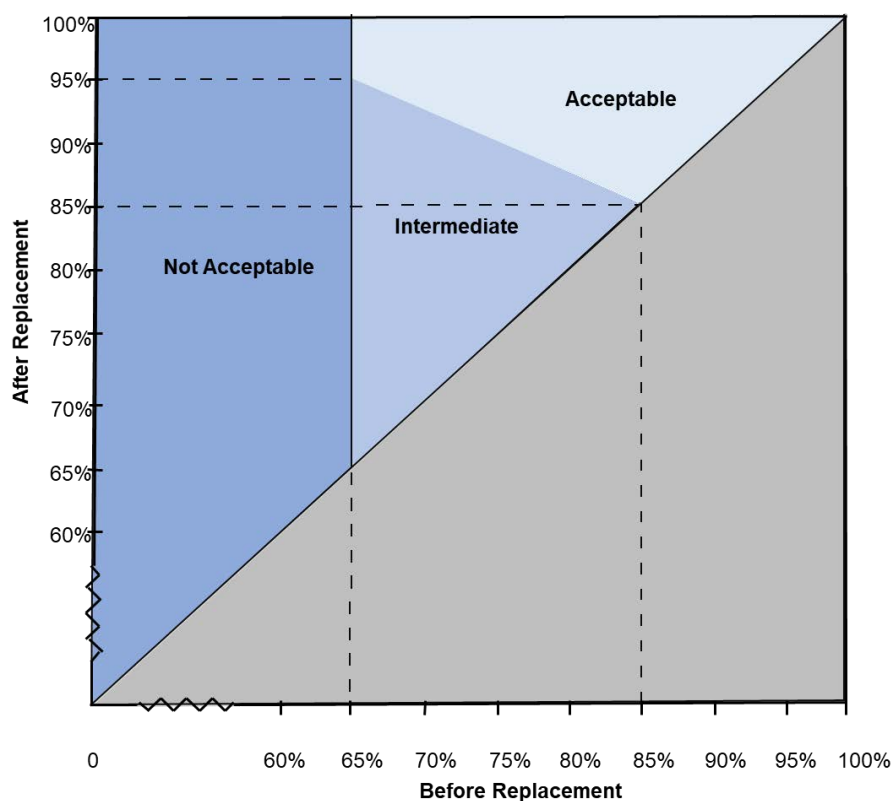
NPMs were strongly encouraged to identify available variables to use for defining the explicit and implicit strata for schools to reduce the sampling variance. See the section “Stratification”, further on in this chapter for more details.

For countries/economies that had larger than anticipated sampling variances associated with their estimates in PISA 2018, recommendations were made regarding sample design changes that were expected to help reduce the sampling variances for PISA 2022. These included modifications to stratification variables and increases in the required school sample.

### **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 6.1 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”.

**Figure 6.1. School response rate standards**



Furthermore, a school with a student participation rate below 33% was not considered as a participating school and data from such schools were not considered for analysis. This was a change from 2018 where 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, but data from such schools were included in the database and contributed to the estimates included in the initial PISA international report, and 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 change from 2018 was implemented so that the minimum

of 33% student participation would be the same for the purposes of calculating and documenting response rates and the data inclusion in the database. 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 selected questions from the background questionnaire (see Annex I).

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 33% student participation. However, there were a few schools in many countries/economies 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 a student response below 33%, treating the school as a non-respondent was likely to introduce less bias and error variance than was treating the students as non-respondents. Clearly the cut-off of 33% is arbitrary as one would need extensive studies to try to establish an optimal cut-off empirically. However, 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.

These PISA standards applied to weighted school response rates. The procedures for calculating weighted response rates are presented in Chapter 10. Weighted response rates weight 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/economies most schools contributed approximately equal weights. Therefore, the weighted and unweighted school response rates were similar. Exceptions could occur in countries/economies 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 13 and 16.

### ***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 a participant. The overall student response rate was computed using only students from schools with at least a 33% 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 survey 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 that could lead to an unapproved definition of the national target population.

Every NPM was required to define and describe their country/economy'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 required approval 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/economy's school sampling frame. Exclusions were often proposed for practical reasons such as unreasonable increased survey costs or complexity in the sample design and/or difficult testing conditions. These difficulties were generally addressed by modifying the sample design to reduce the number of such schools selected rather than to exclude them. 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/economies were instructed to include the schools but to administer the PISA *Une Heure* (UH) form, consisting of a subset of the PISA assessment items, deemed more suitable for students with special needs. Sixteen countries/economies used the UH booklet for PISA 2022.

Within participating schools, all PISA-eligible students 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/economy 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/economy are documented in Chapter 13.

### ***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* set of documents 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 in grade 7 or higher, 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 and to use most recent information available.

All countries/economies used school-level sampling frames as their first stage of sample selection. The *School Sampling Preparation Manual* set of documents indicated that the quality of sampling frames 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 two-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/economies 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* set of documents 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/economies 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 (the latter type of information was not needed by contractors—only by NPMs, thus there was no requirement for contractors to have this type of information on the school frame submitted by NPMs.)
- coded information about the school, such as region of country/economy, school type and extent of urbanisation, which would be used as stratification variables.

### **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 different strata;
- ensure all parts of a population were included in the sample; and
- 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/economy. Implicit stratification consists essentially of sorting the schools within each explicit stratum using a set of designated implicit stratification variables. Examples of implicit stratification variables could be type of school, urbanisation, school size, or minority composition. Implicit stratification, with systematic sampling, is a way of ensuring a 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<sup>[1]</sup>). Guidelines on choosing stratification variables that would possibly improve the sampling were provided in the *Sampling in PISA manual* (OECD, 2016<sup>[2]</sup>).

Annex Table 6.A.2 provides the explicit stratification variables used by each country/economy, as well as the number of explicit strata found within each country/economy. For example, Australia had eight explicit strata using states/territories which were then further delineated by three school types (known as sectors). Australia 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 Annex Table 6.A.2. Annex Table 6.A.2.

As the sampling frame was always sorted by school size within each stratum, school size was always implicit stratification variable, though it is not listed in Annex Table 6.A.2. 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.

### ***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 the treatment of small schools, 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.

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/economies) was equivalent to drawing a simple random sample of small schools. That is, each small school 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/economy 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. Similarly, if only three schools existed in any explicit stratum, instead of taking only two, all three were selected, to increase the efficiency of the sample design.

#### *Sorting the sampling frame*

The *School Sampling Preparation Manual* set of documents 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 first selection number was used to identify the first sampled school in the specified explicit stratum, as described in the section “Identifying the sampled schools”.
- 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, using 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 6.1. Illustration of probability proportional to size (PPS) sampling

To illustrate these steps, suppose that in an explicit stratum in a participant country/economy, 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
0001	550	550	226	Selected
0002	364	914		
0003	60	974	926	Selected
0004	93	1 067		
0005	88	1 155		
0006	200	1 355		
0007	750	2 105	1 626	Selected
0008	72	2 177		
0009	107	2 284		
0010	342	2 626	2 326	Selected
0011	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  $R1$ , while the school immediately preceding the sampled school was designated as the second replacement and labelled  $R2$ . The *School Sampling Preparation Manual* set of documents noted that in small countries/economies, 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 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.

Variations 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, four-digit school code sequentially numbered starting with one within each explicit stratum (each explicit stratum 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 0001 to 0150. First replacement schools in the main survey are assigned the school identifier of their corresponding sampled schools, incremented by 1000. For example, the first replacement school for sampled school 0023 is assigned school identifier 1023. Second replacement schools in the main survey are assigned the school identifier of their corresponding sampled schools, but incremented by 2000. For example, the second replacement school for sampled school 0136 took the school identifier 2136.

### *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 (the first replacement was contacted first, followed by the second replacement if needed). 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 33% 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 33% 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* equal to or above the *TCS* (42 students in most countries/economies). A moderately small school had an *ENR* in the range of one-half the *TCS* to *TCS* (21 to 41 students in most countries/economies). A very small school had an *ENR* less than one-half the *TCS* (20 students or fewer in most countries/economies). 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 yielding 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 proportionally increasing the number of large schools to sample. To determine whether very small schools should be under-sampled 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 under-sampled and the school sample size increased, sufficiently 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 schools that are the very smallest schools ( $ENR$  of 0, 1, or 2) was 20 percent or more of total schools on the frame, and the percentage of students in moderately small schools ( $TCS/2 < ENR < TCS$ ) was 4 percent or more, then very small schools were under-sampled and the school sample size increased.
- If the percentage of students in very small schools ( $ENR < TCS/2$ ) was LESS than 1 percent, and the percentage of schools that are the very smallest schools ( $ENR$  of 0, 1, or 2) was LESS than 20 percent of total schools on the frame, and the percentage of students in moderately small schools ( $TCS/2 < ENR < TCS$ ) was 4 percent or more, then there was no under-sampling of very small schools needed but the school sample size was increased.
- If the percentage of students in very small schools ( $ENR < TCS/2$ ) was less than 1 percent, and the percentage of schools that are the very smallest schools ( $ENR$  of 0, 1, or 2) was 20 percent or more of total schools on the frame, and the percentage of students in moderately small schools ( $TCS/2 < ENR < TCS$ ) was less than 4 percent, then very small schools were under-sampled and the school sample size may have needed to be increased, with the extent to be determined.

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 under-sampling of very small schools was needed nor an increase to the school sample size to compensate for small schools.

The condition included in the second, third, and fourth points above, where the percentage of schools on the frame that are the very smallest ( $ENR$  of 0, 1, or 2) is 20 percent or more, was added in the PISA 2015 cycle and also applied in 2018 and 2022. This modification from earlier cycles 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 under-sampling was otherwise required based on the percentage of enrolment in very small schools, very small schools were under-sampled to avoid having too many of these in the school sample. Even though under-sampling can reduce the number of these in the sample from what could be expected without under-sampling, 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/economies.

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/economy, 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 *P2* will give an initial sample size of approximately  $14 \times 12 = 168$  students, and very small schools of type *P1* 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, *P1* or *P2*, 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 survey overlap control***

Within a given country/economy the main survey for PISA 2022 could occur at approximately the same time as another survey of schools. Because of the potential for increased burden, an overlap control procedure for school sampling was offered. This was used for one country/economy, Norway (to avoid overlap with the ICCS 2022 sample)<sup>1</sup>. This overlap control procedure for each country/economy required that the same school identifiers be used on the PISA and the other study school frames for the schools in common.

PISA implements the sample overlap control procedure in cases where the other study sample is selected before the PISA sample. Thus, for a country/economy requesting overlap control, the national study centre supplied the international contractor with their school frame, 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. Norway did not request this for PISA 2022.

To control overlap of schools between PISA and another sample, the sample selection of schools for PISA adopted a modification of an approach described by Keyfitz (1951<sub>[31]</sub>) based on Bayes' Theorem. To use PISA and ICCS 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, using Norway and overlap with the ICCS as examples for brevity:

$$C_{PROB} = \begin{cases} \max \left[ 0, \left( \frac{PROBI + PROBP - 1}{PROBI} \right) \right] & \text{if the school was a ICCS school} \\ \min \left[ 1, \frac{PROBP}{(1 - PROBI)} \right] & \text{if the school was not a ICCS school} \\ PROBP & \text{if the school was not a ICCS eligible school} \end{cases} \quad \text{Formula 6.1}$$

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

$$CMOS = C_{PROB} \times \text{stratum sampling interval}$$

The PISA school sample was then selected using the line numbers created as usual, as described in an earlier section of this chapter, 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 2022 Technical Standard 1.16 (see Annex I) states that, as in the previous cycles, the international contractor should select the school samples unless otherwise agreed upon. 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/economies' school samples were selected by, and checked in detail by, the international contractor. To enable this, all countries/economies were required to submit sampling information on forms associated with the following various activities and Sampling Tasks (STs) described in Annex Table 6.A.3

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/economies could use for tracking school participation). Annex Table 6.A.3 provides a comprehensive summary of the information required for each sampling task and the timetables (which depended on national assessment periods). Sampling Tasks are also described in detail in further sections of this chapter.

Once received from each participating country/economy, each set of information was reviewed and feedback was provided to the country/economy. 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/economies 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/economies submitted their main survey sampling tasks, the international contractor verified all special situations known in each participating country/economy. 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 administered; the Creative Thinking assessment was being omitted; overlap control procedures with a national or international (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/economies with fewer or only slightly over their target number of assessed students in PISA 2018 had increased school sample sizes discussed and agreed upon. Additionally,



countries/economies which had too many PISA 2018 exclusions were warned about not being able to exclude any schools in the field for PISA 2022. Finally, any countries/economies with effective student sample sizes less than 400 in PISA 2018 also had increased school sample sizes discussed and agreed upon.

## **Sampling Tasks**

### **School samples**

The school sampling procedure was carried out according to the completion of a series of tasks. During each of these tasks, several checks were performed with the data to ensure the quality of the resulting sample. These sampling tasks are the following:

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

- Language distributions were compared with those of PISA 2018 for countries/economies which had participated in PISA 2018. 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 56-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/economy 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/economy's choice of explicit stratification variables. If a country/economy 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/economy 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.<sup>2</sup>

- 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. Therefore, in the case of overlap control with a non-PISA survey, as many as possible implicit stratification variables should become explicit stratification variables.
- If grade or other national option sampling, or special oversampling of subpopulations of PISA students were chosen as national options, checks were done to ensure that each explicit stratum had only one student sampling method applied.

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

- The total national number of 15-year-olds 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.
- For countries/economies 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 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 2018 and previous cycles. Differences were queried.
- The number and percentage of students to be excluded at the school level were checked and the percentage was checked to confirm that it was less than the guideline maximum allowed for such exclusions.
- 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/economy was queried for reasonableness about what the “other” category represented. If it was known the country/economy 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/economies 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 would 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/economies 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 (not applicable for PISA 2022 as there were no three-stage designs). For countries/economies 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*

- The type of school-level enrolment estimate, and the year of data availability were assessed for reasonableness.
- Countries/economies were asked to provide information for each of various school types, whether those schools were included on or excluded from the sampling frame, or the country/economy did not have any 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 2018 as well as previous cycles. Differences were queried.
- NPMs were queried about whether 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

national education authorities, 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/economy'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/economy 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 expected number of assessed students was checked to ensure that at least 6 300 assessed students in CBA countries/economies, and 5 250 assessed students in PBA countries/economies, were expected.
- 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 survey overlap control if applicable.

- 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* set of documents were adhered to.

#### *Sampling Task 11a/b: Reviewing and agreeing to the sampling forms*

- The forms for Sampling Tasks 11a/b were prepared as part of the sample selection process. After the international contractor verified that all entries were correct, NPMs had to perform the same checks and to agree to the content in these forms as quickly as possible.

#### *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 larger than expected based on sampling frame enrolment information. Such situations were queried.
- Large differences in overall grade and gender distributions compared to unweighted 2015 and 2018 data were queried.
- Uneven distributions of student birth months were queried when such distributions differed from unweighted 2015 and 2018 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.

## Student samples

Student sampling was undertaken using the international contractor software, ACER Maple, 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 the 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.

Two countries (Germany and Iceland) 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 two grade 9 classes within PISA-sampled non-SEN schools (except for vocational schools) and all grade 9 classes within PISA-sampled SEN schools

that had this grade. Prior to PISA 2015, Germany assessed all the class-sampled students. For PISA 2022, similar to PISA 2018, to reduce the number of students needing to be assessed for their grade sample from the sampled classes, Germany randomly subsampled 15 students in each sampled class only to participate; the non-selected students in each sampled class were dropped in weighting after applying a ratio adjustment to student base weight for sub-sampled students within each sampled class.

- Iceland had a school census and a student census of PISA-eligible students, as well as a census of grade 10 students.

Two countries (Denmark and France) 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/economy) 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). This form was referred to as a student listing form. The following were considered important:

- Age-eligible students were all students born in 2006 (or the appropriate 12-month age span agreed upon for the participating country/economy). With additional grade samples, including all 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 within each school, using systematic sampling, where the lists of students were first sorted by grade and gender. NPMs were required to use ACER Maple, to select the student samples unless otherwise agreed upon. For PISA 2022, all countries/economies used ACER Maple. The same procedures were used to select the student samples for the Field Trial.

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

PISA was a timed assessment administered in the instructional language(s) of each participating country/economy 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 guidelines 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 8) 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 an impediment 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 survey to be comparable within and across participating countries/economies. 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 I).

## **Teacher samples**

For PISA 2022, as in PISA 2018, a limited number of countries/economies elected to participate in 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/economy. Therefore, the TQ focused on the grade level that most 15-year-old students in the country/economy attend, or in other words, the national modal grade for 15-year-old students. If an adjacent grade level was attended by 30% or more of 15-year-old students in the country/economy, 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 an intact class in a classroom, in small groups in a resource room or one-to-one inside or outside regular classrooms.” Sampling for teachers included all teachers who were currently teaching the modal grade.

Teachers were listed and sampled in ACER Maple as either part of Population ID 1 (mathematics teachers) or Population ID 2 (teachers of other subjects). The distinction between Population IDs 1 and 2 is determined by the meaning of mathematics. Mathematics lessons are the lessons in which algebra, geometry, trigonometry, pre-calculus, and calculus are taught in a curriculum as separate mathematics subjects or taught within a single ‘integrated mathematics’ subject, according to the national/state curriculum. Teachers who teach mathematics lessons were included in Population ID 1, while other eligible teachers are included in Population ID 2.

Ten mathematics teachers were sampled in schools having at least that many listed, or all such teachers, if there were fewer than 10. Fifteen teachers of other subjects were sampled in schools having at least that many listed, or all such teachers, if there were fewer than 15. Within each teacher population (mathematics and non-mathematics), simple random samples of teachers were selected.

## Definition of school

Although the definition of a “school” is not always straight forward and uniform across all countries/economies, 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/economies.

There are exceptions to this, such as when school shifts are more like separate schools than part of the same overall school. However, in some countries/economies with school shifts, this is not the case, and therefore whole schools are used as the primary sampling unit. Similarly, many countries/economies 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 might 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, if differential sampling of certain tracks or programmes is desired, etc.).

Sampling units to be used on school-level frames were discussed with each country/economy before the field trial. Table 6.3 presents the comments from NPMs, in cases where “school” was not the unit of sampling. Where the Sampling Unit column indicates School, this means that the school was the sampling unit. Where it shows Other then something else was used, as described in the comments Annex Table 6.A.4 shows the extent to which countries/economies do not select schools in PISA, but rather something else.

## References

- Jaeger, R. (1984), *Sampling in Education and the Social Sciences*, Longman, New York. [1]
- Keyfitz, N. (1951), “Sampling with probabilities proportional to size”, *Journal of the American Statistical Association*, Vol. 46, pp. 105-109. [3]
- OECD (2016), *Sampling in PISA*, OECD and Westat, [2]  
<http://www.oecd.org/pisa/pisaproducts/SAMPLING-IN-PISA.pdf>.

## Notes

- 
1. The International Civic and Citizenship Education Study (ICCS) is an international comparative study collecting data on democracy and civic education from students around 14 years of age, teachers and school leaders from a representative sample of schools.
  2. ISCED stands for International Standard Classification of Education, an international statistical framework for organising information related to education systems.



## Annex 6.A. Sample design

**Annex Table 6.A.1. Chapter 6: Sampling**

Tables	Title
Table 6.A.2	Stratification variables used in PISA 2022
Table 6.A.3	Schedule of school sampling activities
Table 6.A.4	Sampling frame units

**Annex Table 6.A.2. Stratification variables used in PISA 2022**

Country/Economy	Explicit stratification variables	Number of explicit strata	Implicit stratification variables
Albania	Locations (2); Geographical division (3); Funding (2); Certainty selections	12	ISCED level (3), Gender (5)
Argentina	Region (10); Sector (2); Certainty selections	21	Department (19); Location (2); Level (8); Performance (5)
Australia	State/Territory (8); Sector (3); Certainty selections	25	Geographic Location (3); School gender composition (3); School socioeconomic level (11); ISCED level (3)
Austria	Programme (17); Certainty selections	18	Region (9); Percentage of girls (5); Programme for Statut schools (3)
Baku (Azerbaijan)	Urbanicity (2); Language (2); Status/Funding (2); Certainty selections	5	None
Belgium	Region (3); Form of education – Flanders (5), French Community (3), German Community (2); Funding – for Flanders only (3); ISCED level (4), Educational tracks – for French Community only (4)	31	Type of school – for French Community only (5); Grade repetition (6); Percentage of girls (5)
Brazil	Region (5); Public/Private (4)	20	State (27); ISCED level (5); Urbanisation (2); Capital/Country (2); IDH Quintiles (5); School gender composition (3)
Brunei	School Governance (4); School Composition (3);	7	Sixth Form (3); District (4)
Bulgaria	Type of location (3)	3	Type of school (3)
Cambodia	Location (2); School Type (3); School Zones (5)	18	School management (2); Shifts (2)
Canada	Province (10); Language (2); School size (4); Certainty selections	67	Urbanicity (2); Funding (2); ISCED level (3)
Chile	School Type (4); School level (3); School track (4);	14	School Type (4); National test score level (4); Percentage of girls (6); Urbanicity (2); Geographic zone (4)
Chinese Taipei	School type (6); Location (3); Certainty selections	19	Funding (2); Region (6); School gender composition (3); Municipality (2); Shift offerings (2)
Colombia	Region (2); Urbanicity/School Type (3)	6	Regional entities (96); Main shift (2); School gender composition (5)
Costa Rica	School groups (5)	6	Zone (2); Track (2); Shift (2); Education regions (27); ISCED level (3)
Croatia	Dominant programme type (6); Certainty selections	7	Region (6); School gender composition (3)
Cyprus	ISCED level (3); ISCED programme orientation (3); Funding (2);	8	Urbanisation (2); Language (2)
Czech Republic	School Type (6); Region for school types 1 and 2 (14)	32	Region for school types 3, 4, 5 (14); Gender (3)

Denmark	Immigrant levels and Faroes (5); Certainty selections	6	School type (7); ISCED level (3); Urbanisation (5); Region (5); FO group (3)
Dominican Republic	Funding (2); Urbanisation (2); ISCED level (3)	10	Shift (6); School size (4); Programme (4)
El Salvador	Departamento (14); Location (2);	28	Founding (2); ISCED level (3); Study Commitment (3)
Estonia	Language (3); Certainty selections	4	School type (3); Urbanicity (2); County (15); Funding (2)
Finland	Region (5); Urbanisation (2); Immigrant cluster (6); Certainty selections	30	Immigrant cluster (6); Regional state administrative agencies – for major regions of Northern & Eastern Finland and Swedish-speaking regions only (7); School type (5)
France	Territoire (4); Type (4); Taille (3)	22	Secteur (2)
Georgia	Urbanicity (5); Ownership (2)	9	Language (9)
Germany	School category (3); State – for normal schools only (16)	18	State for SEN and vocational schools only (16); School type – for normal schools only (6)
Greece	Urbanisation (3)	3	Funding and region (15); School type (4)
Guatemala	Urbanicity (2); Funding (4); Certainty selections	9	ISCED (2); Modality of teaching (4)
Hong Kong (China)	School type (5)	5	Student academic intake (4); School gender composition (3)
Hungary	School type (6)	6	Geographical region of Hungary (7); Average mathematics performance in the National ABC 2020 (6)
Iceland	Region (6); School size (4)	24	Urbanicity (2)
Indonesia	Region (4)	4	School type (5); Funding (2); Region (8)
Ireland	School sector (3); School Size (3)	9	School gender composition (4); Socioeconomic quartile (4);
Israel	School orientation (12); Certainty selections	13	ISCED level (3); Group size (2); Socio-Economic status (3); Geographic/Administrative District (2)
Italy	Region (7); Study programme (5); Certainty selections	36	IRegion (20); Types of school (2)
Jamaica	Regions (8); Urbanicity (3); Certainty selections	15	Gender (3); School types (5)
Japan	Funding (2); Orientation (2)	4	Levels of proportion of students taking university/college entrance exams (4)
Jordan	School type / Funding (7); Certainty selections	8	Region (3); Urbanisation (2); School gender composition (3); Level (2); Shift (2)
Kazakhstan	School type (2); Region (17); Certainty selections	19	ISCED Level (2); Location (2); Language (3); Funding (2); Shifts (2)
Korea	School level (3); Orientation (2); Certainty selections	6	Urbanisation (3); School gender composition (3)
Kosovo	Region (7); Certainty selections (Large schools)	8	Urbanisation (2); ISCED (3)
Latvia	Urbanisation (4)	4	School type/level (4)
Lithuania	School language (5); School location – for Lithuanian language (4), for other languages (1); School type – for Lithuanian language (4), for other languages (1); Certainty selections	21	School language 2 (4); School location (5); School type (5); School type 2 (2)
Macao (China)	School type (3); Study programme (2); Language (5)	10	School gender composition (3); Secular or religious (2)
Malaysia	School category (9); Certainty selections	10	School type (18); Location (2); Gender (3); ISCED level (2)
Malta	School orientation/management (3);	3	None
Mexico	School level (2); School type funding(2); School size (3)	12	School program (8); Urbanisation (2)
Mongolia	Location (6); Settlement (4); Certainty selections	16	Property type (3); ISCED orientation (2); ISCED level (3)
Montenegro	Programme (4); Region (3)	12	School gender composition (3)
Morocco	Region (12)	12	Milieu (2); Type (2)
Netherlands	School track (10)	10	None

New Zealand	School size (3); Certainty selections	4	School decile (4); School authority (2); School gender composition (3); Urbanicity (2)
North Macedonia	Language (3); ISCED programme (3)	9	Urbanisation (2)
Norway	School type (2)	2	None
Palestinian Authority	Authority (2); Interventions (3); Certainty selections	7	Region (2); Gender (3); District (25)
Panama	Sub-system of education (3); Urbanicity (2); Funding (2); Certainty selections	16	Educational region (16); ISCED level (3); Programme orientation (4); Language of test (3)
Paraguay	School sector (3); School area (2); School size (3); Certainty selections	19	Region (5)
Peru	Funding (2); Urbanisation (2)	4	Region (26); School gender composition (3); School type (4)
Philippines	Administrative Region (16)	16	School Management (2); Type of Community (3); ISCED Level (3); Gender Composition (5)
Poland	School type (4)	4	Private/Public (2); Locality size (4); School gender composition (3)
Portugal	Geographic region (25); Certainty selections	26	ISCED (3); Funding (2); Urbanisation (3); Curriculum (3)
Qatar	School type (4)	4	Level (5); School gender composition (3); Language (2); Programme orientation (3)
Republic of Moldova	Language (3); Urbanisation (3); ISCED level (3); Certainty selections	28	Funding (2); Study programme (6)
Romania	Programme- ISCED Level (2); Language (3)	6	School location area (2); Development regions (8)
Saudi Arabia	School type (3); Gender (2); Region (5)	30	District (47); School level (2)
Serbia	School type primary (2); Region - for non-primary schools only (5), for primary schools (1); School type - for non-primary schools only (4), for primary schools (1); Certainty selections	22	Region implicit (5); School type implicit (7); Language (2)
Singapore	Public/Private (2); School level (2); Certainty selections	4	School Gender composition (3)
Slovak Republic	School type (3); Region (8)	24	T9 - Three-year average of scores in national testing in math and Slovak (Hungarian) language (7); School type (6); Language (3); Funding (3)
Slovenia	Programme/Level (7)	7	Location/Urbanisation (5); School gender composition (3)
Spain	Region (19); Funding (2); Linguistic model – for the Basque region only (2); Certainty selections	40	Linguistic model - for Basque Country only (3), other regions (1)
Sweden	Funding (2); ISCED level (3); Urbanisation for lower secondary only (3)	8	Geographic LAN – for upper secondary only (21); Responsible authority – for upper secondary only (3); Level of immigrants (3); Income Quartiles – for lower secondary/mixed only (4)
Switzerland	Language (3); ISCED level (3); Urbanisation (2)	15	Sponsorship (2); School type (41); Canton (26); Foreign speaking student share (3)
Thailand	Educational administration (7); ISCED level (3); Certainty selections	15	Public/Private (2); Region (9); Urbanisation (2); School gender composition (3)
Turkey	School Type by Percentile of Performance (36)	36	Statistical Region Unit (12); Location (2); Gender (3)
Ukraine (18 of 27 Regions)	Urbanicity (2); Region (25)	49	ISCED Orientation (3); Language (3)
United Arab Emirates	Emirate (7); Funding (2); Curriculum (5)	47	School gender composition (3); Language (3); ISCED level (3); ISCED programme orientation (2)
United Kingdom (excl. Scotland)	Country (3); School type (6); Region (13), Certainty selections	34	School gender composition (3); School performance – England (6) and Wales (5)

			only; Local authority (7)
United Kingdom (Scotland)	Funding type (3); School attainment (6)	8	Gender (3); Area type (6)
United States of America	Region (4); Funding (2)	8	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 (4)
Uzbekistan	Region (14); Urbanicity (2)	27	Specialization (2)
Viet Nam	Zone (3); Funding (2); Location (3)	15	Region (6); Province (63); School type (4); Study commitment (2)

### Annex Table 6.A.3. 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 ACER Maple	Sampling Task 11a – reviewing and agreeing to the sampling form containing sample design specifics for ACER Maple	Countries/economies had one week to agree to their Sampling Task 11a after TCS was finalized
Review and agree to the sampling form required as input to ACER Maple	Sampling Task 11b – reviewing and agreeing to the sampling form containing records for all of the sampled original and replacement schools and within-school sampling information for ACER Maple	Countries/economies had one week to agree to their Sampling Task 11b after Sampling Tasks 10 and 11a were approved
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

Annex Table 6.A.4. Sampling frame units

Country/Jurisdiction	Sampling unit school/other	Sampling frame units comment
Albania	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
Baku (Azerbaijan)	School	
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	
Brunei	School	
Bulgaria	School	
Cambodia	School	
Canada	School	
Chile	School	
Chinese Taipei	School	
Colombia	Other	"Sedes," or physical location
Costa Rica	School	
Croatia	School	
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	
El Salvador	School	
Estonia	School	
Finland	School	
France	School	
Georgia	School	
Germany	School	Exceptions in SEN schools
Greece	School	
Guatemala	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	
Jamaica	School	
Japan	Other	Programme
Jordan	School	
Kazakhstan	School	
Korea	School	
Kosovo	School	
Latvia	School	
Lithuania	School	If schools have a main building in one place and another building located in a different area, those separate buildings are listed as separate frame units, and if schools do not have that situation, the whole schools are used as frame units.
Macao (China)	School	

Malaysia	School	
Malta	School	
Mexico	School	
Mongolia	School	
Montenegro	School	
Morocco	School	
Netherlands	Other	Locations of (parts of) schools, often parts of a larger managerial unit
New Zealand	School	
North Macedonia	School	
Norway	School	
Palestinian Authority	School	
Panama	School	
Paraguay	School	
Peru	School	
Philippines	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
Qatar	School	
Republic of Moldova	School	
Romania	Other	School programmes
Saudi Arabia	Other	Some schools have two units such SEN programs and regular programs
Serbia	School	
Singapore	School	For public schools, sampling units were whole schools. For private schools, different campuses of private schools were treated as separate sampling units.
Slovak Republic	School	There is type of school, which has the name United school: one individual school with 2 organisation units. Each of the organisation units is separate.
Slovenia	Other	Study programme within ISCED3 schools and whole ISCED2 schools
Spain	School	
Sweden	Other	"School units", some schools have been divided horizontally or vertically so that each part has only one principal
Switzerland	School	
Thailand	School	
Turkey	School	Level of organisation in Multi Programme Anatolian High Schools will be at programme level and the whole school.
Ukraine (18 of 27 Regions)	School	
United Arab Emirates	Other	Separate curricula and also by gender. Whole schools sometimes.
United Kingdom (excl. Scotland)	School	
United Kingdom (Scotland)	School	
United States of America	School	
Uruguay	Other	Night shift is considered a different school
Uzbekistan	School	
Viet Nam	School	

# 7 Translation and Verification of the Survey Material

## Introduction

This chapter describes the translation and adaptation procedures, the linguistic quality control (verification) procedures for both paper-based (PB) and computer-based (CB) materials in PISA 2022, as well as the upstream linguistic quality assurance procedures used to produce the source versions of the PISA instruments.

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. To achieve this, strict procedures for the localisation (adaptation, translation, and validation) of national versions of all survey instrumentation were implemented in PISA 2022, as in all previous PISA rounds.

These procedures included upstream and downstream linguistic quality assurance processes, further explained below.

Upstream Linguistic Quality Assurance Processes include the following aspects:

- Optimisation of 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).
- Implementation of a double translation design with a final reconciliation.
- 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.
- Production of item-by-item translation and adaptation notes.
- Training of national staff in charge of the translation/adaptation of the instruments.
- Centralised trend material transfer.

Downstream Linguistic Quality Control Processes include the following aspects:

- 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.
- Centralised management of the changes and updates in the trend materials.

## PISA Countries/economies, Languages, Scope and Verifier training

The countries or economies participating in PISA 2022, referred to in this report as PISA Countries/economies, were responsible for the translation and adaptation of their instruments. Annex Table 7.A.2. lists the verified language versions with the following additional information:

- ISO (three-letters) Code 3366
- The last cycle in which they participated in PISA
- The mode of administration (PB for Paper Based or CB for Computer Based assessment)
- The change of mode compared to the last cycle (PBA → CBA)
- Whether the version was adapted from the English or French source, from the common base version in Spanish or Chinese, or from borrowed version from another country/economy
- The international options that underwent the verification process: CT (Creative Thinking), FL (Financial Literacy), ICQ (Information and Computer Literacy Questionnaire), TCQ (Teacher Questionnaire), WBQ (Well-being Questionnaire), UH (Une-heure test and questionnaires), PAQ (Parent Questionnaires).

While most of the PISA 2022 Countries/economies has also administered the assessment in PISA 2018, five Countries/economies with six versions were new to PISA 2022: El Salvador, India with two languages, Hindi and English, Jamaica, Mongolia and Uzbekistan. In total, 113 language versions in 54 languages for 86 PISA Countries/economies were verified in PISA 2022. The table does not include minority language versions that represented less than 10% of the target population and were not centrally verified.

## Materials subject to verification

The following materials were subject to international verification before the Field Trial:

### **Cognitive units**

The PISA 2022 cognitive assessment consisted of the units from the three core domains, compulsory for all the PISA Countries/economies, and the international options. These include the following units:

### **New Mathematics units**

Mathematics was the main domain in PISA 2022: 61 of newly developed Mathematics units were translated and verified in three batches. In past cycles, the PISA Countries/economies administered one of the two “easy” or “hard” clusters. Both clusters were administered by all the PISA Countries/economies in PISA 2022, referred to as cluster 6A and 6B. Either 6A or 6B cluster was verified as new for all the Countries/economies. New Mathematics units were computer-delivered and were translated and verified in XLIFF format (tagged XML Localization Interchange File Format) in the open-source CAT (computer-assisted translation) tool OmegaT. The units were released in 4 batches for translation and adaptation, with 6A or 6B cluster in a separate batch. See Annex Table 7.A.3. .

From this pool, 16 new units and 10 new items were dropped for the Main Survey. See Annex Table 7.A.4.

### *Financial Literacy units*

Three new Financial Literacy units were added to the cognitive assessment pool for PISA 2022, translated, verified, and administered by the Countries/economies that have chosen this international option in Field Trial and Main Survey.



### *Creative Thinking units*

Creative Thinking was the new domain introduced in PISA 2022 as an international option, with 21 units. Unit T54, Infographics was not administered in the Main Survey. See Annex Table 7.A.5.

### *Mathematics units (trend)*

45 trend units were administered in the Field Trial, from which 2 units and 3 items were dropped for the Main Survey.

### *Financial Literacy units (trend)*

Three new Financial Literacy units were added to the test pool for PISA 2022, translated, verified, and administered by 21 Countries/economies that have chosen this international option in Field Trial and Main Survey in 28 national versions.

### *Reading units (trend)*

Forty-nine Reading Literacy trend units were administered in the PISA 2022 Field Trial and Main Survey. For the Countries/economies new to PISA, the trend Reading units were translated and verified as 'new' materials following same workflow and procedure as for new Mathematics units.

### *Science units (trend)*

Twenty-four units from the trend Science instruments were administered in PISA 2022 Field Trial and Main Survey. Like Reading units, Countries/economies new to PISA followed the workflow and procedures same as for the new Mathematics units.

### *Orientation, Help, Interface and Test flow files (XYZ files)*

There was one new 'orientation' and one new Help file verified for all CBA countries; orientation file for FL was verified for Countries/economies taking those options. The Creative Thinking orientation file was translated and verified with the CT units.

### *Orientation, Help, Interface and Test flow files (XYZ files) (trend)*

There were nine files with other widgets, or "XYZ files", included interfaces for the calculator and Math editor, generic navigation elements, a help file, the interface for the test environment orientation files for the questionnaires, reading, and test flow. The new PISA Countries/economies that administered the units on computer, translated these files in OmegaT, and they were all verified.

### *Paper-based clusters*

For Countries/economies administering PISA 2022 as a paper-based assessment (PBA Countries/economies), the cognitive test consisted of trend units only, as no new PB items were developed for PISA 2022. For Countries/economies that were new to PISA, all 44 Math, 32 Science, 22 Reading units and 4 Reading components were treated as 'new' materials and underwent the translation and/or adaptation process.

## **Contextual Questionnaires**

There were two required contextual questionnaires, administered by all participating countries, and five optional questionnaires:

### *Required Questionnaires*

- School Questionnaire (SCQ) with 83 questions administered on the Questionnaire Adaptation Tool (QAT) for CBA countries; for PBA countries 69 questions were translated and verified in Main Survey Word format, and administered on paper;
- Student Questionnaire (STQ) for PBA countries was administered in paper-based format (MS Word) in two Booklets, each of them consisting of 15 Core questions, identical between the two booklets, as well as 30 additional questions in Booklet 1 and 42 additional questions in Booklet 2. The CBA countries administered the Student Questionnaire with 168 questions in the QAT.

The Global Crisis Module (GCM) were questions added in SCQ and STQ following the outbreak of the COVID-19 pandemic. Counts of GCM questions are included in the counts above.

### *Optional Questionnaires*

- Parent Questionnaire (PAQ) with 45 questions available in paper-based format for both PBA and CBA countries. The Parent Questionnaire was verified in 13 languages (corresponding to 20 national versions) in 17 countries, all of these CBA countries. No PBA country opted for the Parent Questionnaire.
- Information and Communication Technology Questionnaire (ICTQ) with 14 questions administered in the QAT (70 versions verified for 57 CBA countries);
- Teacher Questionnaire (TQ) with 77 questions included in the QAT (24 versions verified for 20 CBA countries). Some questions were addressed specifically to mathematics teachers.
- Financial Literacy Questionnaire (FLQ) with 14 questions included in the QAT (31 versions for 23 countries) by countries that also opted for the Financial Literacy cognitive assessment.
- Well-Being Questionnaire (WBQ) with 25 questions included in the QAT (21 versions verified for 16 CBA countries).

## **Verifier qualifications, training and instructional materials**

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

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 and/or French into their target language,
- if 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,
- if possible, familiarity with the main domain assessed,
- a good level of computer literacy and experience with computer-aided translation tools (CAT tools),
- if possible, experience as teachers and/or higher education degrees in psychology, sociology, or education.

All verifiers were invited to attend one of the two seminars, based on the verification schedule of their country. In total 32 verifiers of early-testing countries and 10 members of the cApStAn team attended the

first training seminar in June 2019, and 20 verifiers and 10 cApStAn team members the second training seminar in September 2019. A 2-day verifier training seminar was organized by cApStAn in Brussels on 31<sup>st</sup> May and 1<sup>st</sup> June 2019. In total 55 verifiers and 10 members of the cApStAn team attended the seminar. Those verifiers who were not able to come to the seminar were trained through remote Webinars in July and/or August of 2019.

The main aim of the training was to provide verifiers with background information on PISA 2022 in general, and on the verification task in particular. Verifiers were divided into four different groups based on two criteria (experienced/new and full verification/focused verification process) to attend parallel sessions:

- **Experienced verifiers** – verifiers who had participated in previous PISA cycles and had already acquired experience in verifying PISA materials.
- **New verifiers** – verifiers who had been recruited for this cycle of PISA.
- **Verifiers of adapted versions** – verifiers verifying a version adapted from the French or English source version, from the Spanish or Chinese common base version, or from a verified version produced by another National Centre.

Each group participated in three sessions:

- **Cognitive Materials** – Topics for this session included: nature and new features of the new Mathematical literacy units, challenges of mathematics units compared to other domains; structure of the TAS (Test Adaptation Spreadsheet), as well as the overall verification workflow using the portal previews. The session included hands-on exercises where verifiers edited mock XLIFF files using OmegaT, previewed the resulting file on the PISA portal and documented their findings in a TAS, under the supervision of the cApStAn trainers. The session for new verifiers' group included a generic part explaining the essence of the verification task and more background information on the PISA survey, while this was omitted in the presentation for experienced PISA verifiers. Similarly, the session for verifiers of adapted versions focused on what is relevant for this procedure, drawing examples from adapted versions in previous cycles.
- **Questionnaires** – In this session, the differences in procedure and focus of questionnaire verification vs. verification of cognitive materials was explained. There were also hands-on exercises, where verifiers were asked to work in the Questionnaire Adaptation Spreadsheet (QAS) and on OmegaT, and to verify mock translations.
- **Documentation and tools** – This session concentrated on the principles of documenting verification outcomes using the verifier intervention categories (See Annex Table 7.A.2. ) in a way that is informative, concise, and useful to all parties involved. Examples from previous cycles were discussed among the group to illustrate best practices in comment writing.

Tailoring the sessions to smaller groups proved to be effective in the PISA 2015 and PISA 2018, so the same approach guided the organisation of the trainings for PISA 2022.

Day 1 of the seminar was devoted to OmegaT. During the morning plenary session, the CAT tool and its features were introduced. The group was then split in parallel sessions to give the verifiers the opportunity to perform some practical exercises in smaller groups. A specific meeting for verifiers of right-to-left languages was also organised. At the end of the day, the groups were reunited for a general question-and-answer session.

Day 2 included the following sessions:

- **General PISA session** – Overview of the PISA 2022 Field Trial.
- **Cognitive Materials** – Topics for this session included: a generic part explaining the essence of the verification task and more background information on the PISA cognitive materials, the overall verification workflow, the nature and challenges of New Maths and Creative Thinking units. For the translated versions, the verifiers were divided in two smaller groups. The session for verifiers of

adapted versions focused on what is relevant for the versions adapted from one of the source versions, from a common base version or from a translation borrowed from another country.

- **Documentation and tools** – This session concentrated on the principles of documenting verification outcomes using the verifier intervention categories in a way that is informative, concise and useful to all parties involved. The novelty of the standardised comments was also illustrated. Some practical exercises were organised.
- **Questionnaires** – In this session, the differences in procedure and focus of questionnaire verification vs. verification of cognitive materials was explained. The questionnaire workflow was presented, and there were also hands-on exercises, where verifiers were asked to work in the Questionnaire Adaptation Spreadsheet (QAS) and on OmegaT, and to verify mock translations.
- **Coding guides** – In this session, the focus of verification of the coding guides was explained and the Countries/economies were explained how to take advantage of the translation memories<sup>1</sup> that are coming from the cognitive units. A few sample responses were shown as example.

Splitting certain sessions in smaller groups and organising hands-on exercises proved to be effective in past cycles, so the same principle was followed for PISA 2022.

## Testing languages and translation/adaptation procedures

National project managers had to identify the testing languages according to the PISA technical standards and following the instructions given in the School Sampling Preparation Manual and to record them in the sampling form Sampling Task 0 (ST0) for agreement by the PISA Contractors.

In addition, based on the approved ST0, and prior to the Field Trial, national project managers had to complete 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.

Annex Table 7.A.6. summarises the Field Trial translation procedures for tests and questionnaires, as described in the confirmed translation plans. The figures in the table do not include minority language versions that represented less than 10% of the target population and were not centrally verified.<sup>2</sup>

The total number of the versions in Annex Table 7.A.6. would not represent the total number of verified versions because some Countries/economies had different procedures for different domains or questionnaires, e.g. Romania double translated the cognitive units from English with cross checks against the French source version, but for the Reading Literacy trend units that were double translated from English and French, Colombia adapted the common reference version but double translated the Parent Questionnaire from English source.

Note that for the Catalan, Galician and Basque versions, the cross-checks were made against the verified Spanish version of Spain rather than against the other source version.

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. In previous survey administrations we found that high-quality translations and high level 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 potentially reduces cross-country differences due to translation differences.

Co-operation between countries sharing the same language was therefore fostered and facilitated. To this effect, workable models were designed so that verified versions from one country could be adapted by another country.

Different scenarios of sharing were applied in the following cases:

- As in previous cycles, the model followed by German-speaking countries was highly efficient: the German version of each of the components of the assessment material was double translated and reconciled by one of the German speaking countries, then verified, and adapted by the other countries who administered that component. The adapted versions were then verified.
- A Spanish common reference version of the new test materials was produced by an independent contractor and shared by the Spanish-speaking countries.
- A Chinese version of the new test materials was produced by an independent contractor and shared by the Chinese-speaking Countries/economies.
- A Russian common reference version was fully verified and then adapted by Azerbaijan (Baku), Estonia, Kazakhstan, Latvia, and Moldova.
- Finally, Bosnia and Herzegovina, Montenegro and Serbia shared the translation effort translating each one part of the assessment and then adapted the verified versions to their local contexts.

## Development of source versions

### ***Translatability assessment***

Translatability assessment is 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 challenges 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 2022 items before the source versions were finalised. This process, referred to as the translatability assessment, was first implemented in PISA 2015.

Translatability assessment consists of submitting draft versions of new items to a pool of experienced linguists covering a broad range of language groups. The linguists 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 that could affect translatability.

For both new Questionnaire items and New Maths and Creative Thinking items there were always at least three linguists from different language groups evaluating each item. The approach was for each linguist to first mentally translate each item allocated to them. When the item appeared straightforward to translate, it was classified as "straightforward." When the linguist found the item somewhat difficult to translate/adapt or identified a potential cultural issue, they went through the exercise of (i) producing a written translation of that item; (ii) selecting the relevant translatability category (such as "Unnecessarily complex" or "Potential cultural issue") – see Annex Table 7.B.1 (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 challenges that translators might face if no pre-emptive action were taken.

The feedback from the different linguists was then collated by a senior linguist at cApStAn and reviewed by the translation referee. The senior linguist 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 particular item. The senior linguist produced a Translatability Report, which was then sent to the item developers for review. The item developers then used the report 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, the aim was 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 PISA, it has been a requirement 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 into a second language. As in previous PISA survey administrations, the English-to-French translation process proved to be very effective in detecting issues not perceived by the item writers, and in anticipating potential problems for translation into other languages. A number of ambiguities or pitfall expressions could be spotted and therefore avoided in the source versions by slightly modifying both the English and French source versions. As a result, the list of aspects requiring national adaptations could be refined; and further translation notes could be added as needed.

The new PISA 2022 items were first drafted in English, and then a parallel source version of the items was produced in French. The parallel source version was produced for the new Mathematical literacy items (stimuli, items, and scoring rubrics for open-ended items), the newly developed items for the School Questionnaire, Student Questionnaire, Information and Communication Technology (ICT) familiarity Questionnaire, as well as the assessment materials for Creative Thinking (stimuli, items, and coding guides). No French source version was produced for the new Financial Literacy items.

The workflow for producing the French source was the same for newly developed PISA 2022 Mathematics units, Creative Thinking units and Questionnaire materials. Once feedback from the translatability report and from country reviews was integrated into the revised units in XLIFF format, the translation monitoring forms in Excel format (Test Translation Spreadsheets, TTS) were prepared for the translation process into French.

There was one TTS for each batch of units and questionnaires. The form was designed to include the whole history of the process and to accommodate (i) comments from translators 1 and 2; (ii) comments from the reconciler (about FRA or about ENG source); (iii) feedback from the domain expert; (iv) consolidated feedback from the lead reconciler (about FRA or about ENG source); (v) first reactions from the test developers, (vi) issues reported during the equivalence and linguistic purity check (ELPC), (vii) second round of feedback from the item developers and (viii) proofreading at the end of the process and potential comments about residual mistakes.

In the TTS, some provisional item-per-item translation and adaption guidelines from the TA were already included for reference and all players were invited to review these and complement with new guidelines as difficulties were identified. The final item-per-item guidelines were then used to populate the Field Trial Verification form.

The translation of the cognitive units for Mathematics and Creative Thinking was done using XLIFFs so consistency could be maximized from the very beginning of the process. The Questionnaires were received in Main Survey Word format. In PISA 2022, OmegaT was also used for the production of the questionnaires

to guarantee the same level of consistency as for the cognitive units. All materials went through a dedicated workflow on the PISA portal.

The workflow was streamlined so that the item-per-item translation and adaptation notes were formulated while the English source version was being finalised. This allowed monitoring the relevance and effectiveness of these notes early on and making necessary adjustments as the parallel source version was produced. The source version optimisation also included work with the Core A Contractor to apply segmentation rules, and to prepare style guides and rule sets for automated consistency checks.

A team of six translators, three reconcilers, two domain experts (one for the Mathematics units and one for the Creative Thinking units and Questionnaires), four equivalence and linguistic purity check reviewers and one proof-reader was set up to produce the PISA 2022 French source. Most members of the team had already participated in producing the French source version of PISA 2018 instruments.

Before the start of the translation, a training workshop with all translators and reconcilers of the parallel source was held in Brussels in December 2018. All translators, reconcilers and domain experts attended the face-to-face training workshop. The training programme included a session on the translation of mathematical language and a hands-on training session to hone the translators' and reconcilers' skills in using specific computer-aided translation tools to their full potential. Sample materials from this cycle and interesting examples from the translatability assessment were used to refresh their memories, and hands-on exercises were organised to introduce the PISA portal and the tools used by cApStAn for this cycle of PISA, including OmegaT, the computer-aided translation tool. There was also an OmegaT helpdesk available throughout the translation process.

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.

The team of translators consisted of one translator who focused primarily on accuracy and systematically conveyed each piece of information in the target version, as well as one translator who concentrated primarily on fluency. As shown in Figure 7.1, the workflow began with producing the two independent translations, T1 and T2. The work was split between Questionnaires translation and cognitive item translation for Mathematics and Creative Thinking. Both translators received the same materials at the same time and delivered their translations to the reconciler on the same date.

**Figure 7.1. Translation workflow for the production of a French source version of newly-developed PISA 2022 Mathematics units**



Both for the new Mathematics units and the questionnaires, translation memories were created from the PISA 2018 and PISA 2015 French source of the questionnaires and added as reference. Translations of the trend questions were thus pre-populated in OmegaT, and all players were instructed to align the translations of the new questions to the trend ones. A glossary of compulsory adaptations, so called “forced

adaptations” from the previous cycles was also prepared and included in the OmegaT projects. Special attention was given to consistency across questionnaires focusing on scales, recurring instructions and forced adaptations. The translation memories from the previous cycles were useful for obtaining better consistency, especially for the questionnaires and the recurring instructions in the new Mathematics units.

The main task of the reconciler was to merge the two independent translations in such a way that the resulting national version is as equivalent as possible to the initial source version while the wording is as fluent as possible. Correspondingly, it was the lead reconciler’s responsibility to finalise their single translation for the coding instructions. In particular, the reconciler ensured consistency between the French version of coding instructions and the French reconciled version of the stimuli and items, and between the English and French source versions. The lead reconciler collated the documentation on all cases where the double translation process (and single translation process for the coding guides) revealed possible flaws in the initial source version and established the communication with the item developers.

The reconciler received the OmegaT packages containing the source XLIFF files, the translation memories from the two translators (T1 and T2) and the Excel monitoring sheet for that batch. The advantage of using XLIFFs already at this stage (instead of Excel files or storyboards) was that it was possible to preview both the English and French version of the unit on the portal, so each translation could be reviewed in its real context. Another important advantage of XLIFFs is that translation of recurring elements could easily be harmonized using the translation memory utility in OmegaT. During this process, the reconciler could enter comments in the Excel monitoring sheet for the attention of the domain experts and the lead reconciler. These comments could relate to the translation and adaptation guidelines, to the English version (linguistic or contents) or to the French version. There were therefore different columns devoted to these comments. The column "Reconciler comment about ENG source" contained reconciler comments about linguistic or content issues as well as some recommendations or suggestions about the ENG wording. These suggestions were mainly aimed to improve consistency or to facilitate the translation into the different PISA languages. Suggestions for item-by-item translation and adaptation notes could also be included in this column. In the column "Reconciler comment about FRA source", the reconciler could explain some of the choices made and document issues for which the domain expert’s advice was requested.

Two domain experts from France reviewed the reconciled translations of the new assessment items from the Mathematical literacy and Creative Thinking domains as well as of the new questionnaire items. The domain experts’ task was to check whether the terminology was deemed appropriate for 15-year-old students; to ensure that the prompts and instructions were clear and relevant, and to evaluate whether, from their expert’s perspective, the cognitive items seemed to measure the same knowledge and skills across the two languages. For the questionnaire items, the domain expert was asked to evaluate that the instruments would collect the same information in each language. The domain experts’ feedback was then processed by the lead reconciler, who either implemented a change directly, or to added it to a compilation of issues that required input from the item developers at Core A and Core B3.

The feedback from the reconciler and the domain experts about the English version was then consolidated by the lead reconciler and shared with the item developers, who reacted to both the reconciler’s and the domain experts’ comments and provided suggestions for edits or in some cases a completely new version of the source wording in English. If a proposed change was relevant for the English master version, the updated English version was entered in the Excel monitoring sheet and the French version was then updated as needed during or after the equivalence and linguistic purity checks.

The interaction between the lead reconciler and the test and questionnaire item developers contributed additionally to the maintenance of semantic, linguistic and insofar as possible, psychometric equivalence between the two parallel source versions. The discussion between the different players was performed by documenting the issues in the TTS. Special attention was given to evaluating the impact of each edit on other parts of the materials and ensuring that the Core A and B3 item developers echo all necessary modifications in the English source.



Once the feedback from the lead reconciler and the item developers was reflected in the French source, it was submitted for a linguistic purity check and semantic equivalence check. These two checks were performed in tandem by (i) a senior staff member of cApStAn who is bilingual English/French and has expertise in the international verification of the PISA materials, who focused primarily on the finer residual equivalence issues; and ii) a native French linguist, who focused primarily on the finer points of strictly correct French language usage. The feedback from this step consisted of comments, suggestions for rewording (sometimes of the English text instead of or in addition to the French text), and proposals for translation/adaptation guidelines.

A senior cApStAn consultant processed the results of the feedback of these two steps simultaneously and shared the reports with the item developers when the reported issues had a potential impact on the English master version. This led to the second round of updates in the English source. Whenever a change in the French version was required, the final version was inserted in a specific column of the monitoring sheet, and this was then centrally transferred into the French XLIFF file by the proof-reader.

Once the item developers' feedback had been implemented, a proof-reader reviewed the final proofs in XLIFF format. The proof-reader saw the materials for the first time in this step. This allowed them to review the final version of the French source version with a 'fresh eye', and correct residual typos, as well as grammar and syntax errors. The proofreader used the 'preview' utility on the PISA portal to proofread the materials. This allowed them to view the items exactly as the respondents would see them. When an issue was spotted, the necessary changes were made in the corresponding XLIFF; then the proofreader would refresh the preview window in order to check that the modifications were correctly implemented. The edits were limited to corrections of outright errors overlooked in the earlier steps or accidentally introduced when processing the feedback from the equivalence and linguistic purity check. The proofreader also left comments in the TTS about any residual issues identified at this step (for instance, incorrect final layout, source updates not implemented etc.) for the item developers' attention.

The coding guides for open ended items were single translated by one of the translators from the team who produced the coding guide for the particular domain, which was first reviewed by the reconciler and the domain expert and then consolidated by the lead reconciler. Finally, the coding guides went through the equivalence and linguistic purity check process and final proofreading.

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 second approach 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). 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 second 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 literal 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 were 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. For the optional Financial Literacy domain, the units were double translated from English only, as there was no French source version of these units.

## Training and instructional materials for national translation teams

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 2019 in Vienna 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 the questionnaires, new cognitive units, trend units and coding guides, separately for the computer-based and paper-based administration, and separately for the new PISA Countries/economies.

### ***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) was required for test and questionnaire materials, but not for manuals, coding guides and other logistic material. In double translation, it was recommended that one independent translator use the English source version while the second use the French version. In countries where the National Project Manager (NPM) has difficulty appointing competent translators from French and English, double translation from English or French only was considered acceptable; in such cases it was 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.

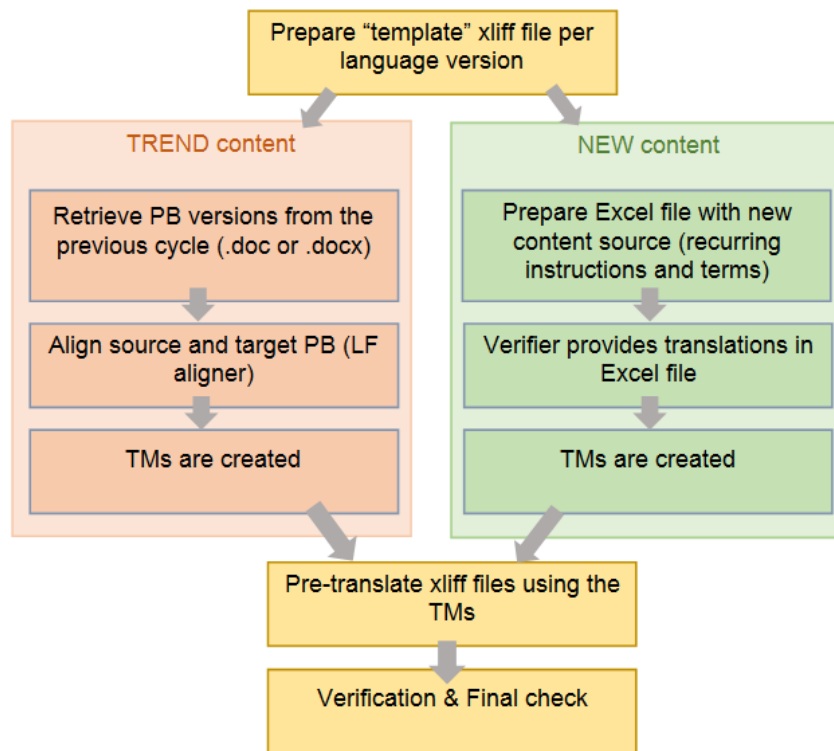
## Centralised trend material transfer

Cognitive units were administered in paper-based format (MS Word) until and including PISA 2012. In PISA 2015, most participating countries switched the mode of administration from PBA to CBA, but there were still some countries that remained with the PBA. In PISA 2022, some of those countries also switched to CBA.

As the trend contents need to remain identical across cycles, the transfer of trend contents from PBA to CBA, i.e. from Word to XLIFF, was centrally managed, as it was in PISA 2015 and PISA 2018. To do this operation, a semi-automated process (different from the more manual process applied in 2018) was adopted. National centres were then asked to review their transferred units using the preview widget on the PISA portal and report any transfer error or residual issues identified in the trend materials using change request forms (in Excel format). Approved changes were then centrally implemented by the contractors.

The workflow of the trend transfer process is shown in Figure 7.2. It details the two parallel workflows that have been developed to transfer the content of the Trend PBA units into the new CBA format. First the PBA materials were extracted from Word and aligned to produce a Translation Memory (TM). Then, the new content that was specific to CBA environment, like specific instructions such as “Click on”, or “Select”, were translated so that these could already be used to pre-translate the CBA xliiffs. Once this pre-translation phase was completed, Quality Assurance checks were performed and translated segments were locked in the OmegaT projects. These transferred materials were then uploaded to the PISA portal for the countries/economies to review. Any residual issue was then documented by the countries/economies and corrected centrally. The countries/economies did not have editing rights to trend content at any stage of the process. This approach prevented 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.

Figure 7.2. Trend Transfer process diagram



## Questionnaire adaptation negotiation

Questionnaire verification before the Field Trial aims to ensure cross-linguistic equivalence of the national versions of the data collection instruments. This process began with the negotiation of national adaptations documented in the Questionnaire Adaptation Spreadsheet, referred to as QAS in this report.

In the questionnaires, national adaptations are defined as intentional deviations from the source, aiming to reflect the national context and to keep the comparability on the international level at the same time. A set of these national adaptations was compulsory, such as country-specific response options in a question that asks about education levels, types of school, or language spoken at home. Beyond these "forced adaptations", countries could propose requests for additional adaptations in the QAS.

Countries proposed their adaptations to new items in the QAS and provided a back translation in English and a justification for the adaptation, as needed. Based on the back translation and the justification, the questionnaire team either agreed to the proposed changes, or asked the National Centre to further adjust the translation to correspond to the source and ensure across-country comparability. This dialogue between the National Centre and the contractors took place in the QAS until an agreement was reached. Then the country-specific "national source" was created by the questionnaire team.

The National Centre implemented the agreed adaptations in their national versions. CBA countries encoded the adaptation directly in the Questionnaire Authoring Tool (QAT).

After having tested the different scenarios (rules and filters) advised by Core A (ETS Data Management), countries uploaded the QAS documenting the negotiation and released the national questionnaires for the next step in the workflow, i.e. verification.

For the first time in PISA 2022 the questionnaire verification was aligned with the cognitive materials in terms of technology, which meant using OmegaT for both. When the negotiation of national adaptations was completed, a national source was created on the Questionnaire Authoring Tool (QAT). The national source was then exported from QAT in XLIFF format for the use in OmegaT. Trend items were centrally populated in OmegaT and locked for editing. The Countries/economies had the possibility to request the changes to trend items within the QAS. These change requests were then negotiated with the Questionnaire Content Team and if agreed, implemented by the verifier during the verification step.

For Countries/economies switching from PBA to CBA, translation memories were created from the PISA 2018 Word files of the questionnaires and transferred on the QAT. The translation of trend questions was thus pre-populated in OmegaT, and all players were instructed to align the translation of the new questions to the existing trend translations.

For PBA versions, the Countries/economies were responsible for maintaining their trend translations.

## International verification of the national versions

As in previous PISA survey administrations an independent team of expert verifiers were appointed and trained by the international contractors to verify each national version against the English and/or French source versions to ensure high-quality standards and assessment materials and contextual questionnaires.

### ***New computer-based test units***

Of the 88 Countries/economies participating in the PISA 2022 Field Trial, 5 participated in the paper-based assessment (PBA). The remaining 83 Countries/economies participated in the computer-based assessment.

Computer-based units were translated and verified using XLIFF files on OmegaT. The files 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 compare the translated segments to the source one by one in OmegaT, while consulting previews on the portal and checking item-specific guidelines and comments from the national centres in the Test Adaptation Spreadsheet (TAS). Where corrections were needed, the verifiers implemented them in OmegaT and documented their interventions in the TAS, using a predefined drop-down menu to assign the change to the appropriate intervention category.

Once a domain was verified, reviewed and finalised on the portal, the translation referee was able to download the TAS annotated by the verifier. The referee would then go through each verifier and country 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 OmegaT packages and TAS 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 7.3 shows a sample test adaptation spreadsheet from the PISA 2022 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, and
- a tool permitting quantitative analysis of verification outcomes.

**Figure 7.3. Sample of a test adaptation spreadsheet (TAS) from the PISA 2022 Field Trial**

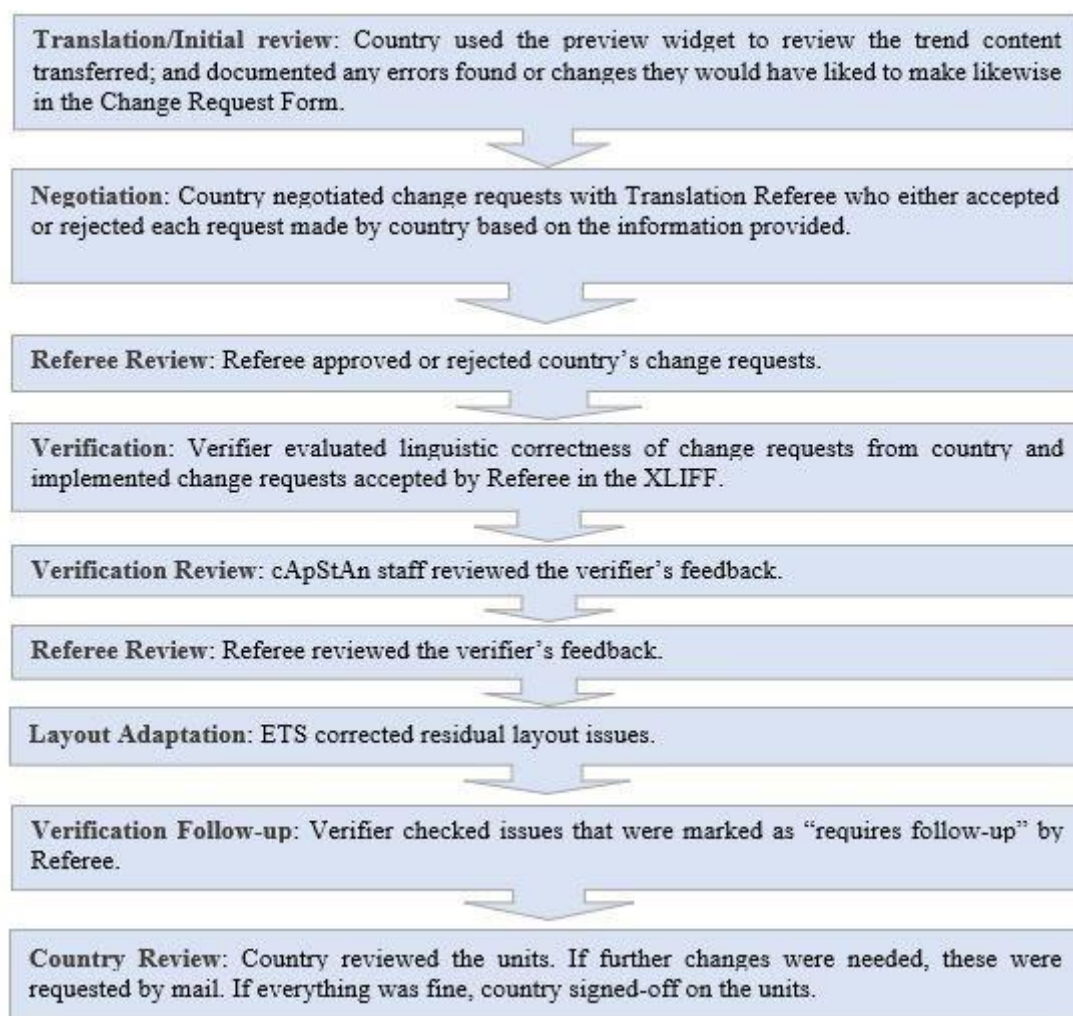
ENGLISH SOURCE VERSION	ITEM-SPECIFIC TRANSLATION/ADAPTATION GUIDELINE	PARTICIPANT COMMENT (ADAPTATIONS, DOUBTS, DIFFICULTIES)	INTERVENTION CATEGORY	VERIFIER COMMENT	Corrected?	TRANSLATION REFEREE COMMENT	CORRECTION STATUS	PARTICIPANT POST-VERIF COMMENT	VERIFIER FINAL CHECK	VERIFIER FINAL CHECK COMMENT
She is correct, because the height of a medium box is 2/3 the height of a large box.	Pattern: response options start in the same way. If possible, reflect the pattern in the target.		OK	T&A guideline followed.						
She is correct, because 3 medium boxes can always be fit into the same space as 2 large boxes.										
She is not correct, because none of the interior storage dimensions of truck A are multiples of 0.75, which is the height of a large box.	Please do not use the translation of the word 'area' in reference to the compartment because this unit is all about volume.		Inconsistency	"box" translated inconsistently within item. T&A guideline followed.	Yes	Please keep the verifier correction	REQUIRES FOLLOW-UP	ok	OK	

### **Cognitive trend units**

For cognitive trend units, i.e. units that the Country/economy has administered in one of the previous cycles, it is essential that the unit is administered in the exact same form to be able measure trends in time. For this reason, centralized trend management was deployed. The Countries/economies did not have editing access to trend units, i.e. units that the Country/economy had administered in one of the previous cycles, at any point of the translation, adaptation, and verification workflow. They were given the opportunity to request changes to trend units, if for example a residual linguistic error or outdated adaptation was identified. The Countries/economies documented these requests with a justification for change in a change request form (Excel file). If the translation referee and the verifier agreed that a change is indeed acceptable, it was implemented by the verifier.

The verification workflow for the trend units is shown in Figure 7.4

Figure 7.4. Verification workflow of trend items



For trend items there was no difference between adapted and translated versions as regards the Change Request Form and the overall procedure.

As a National Centre reviewed the Trend items, it reported any linguistic or content-related request for modification and then submits the annotated Trend Change Request Form to the Translation Referee for approval. All errors related to the trend transfer procedure, that were thus not changes versus trend content, were automatically approved. For any requests that would mean a real content related change versus trend the referee's role was to evaluate whether the requested changes were legitimate or not and could have an impact on the trend data collection. The result of this arbitration process was a Change Request Form where the countries' requests were either approved or rejected by the referee.

The following type of change requests were generally accepted:

- requests to correct outright errors, such as typos, blatant grammar issues, mistranslations,
- requests to correct outdated adaptations, e.g. change of currency,
- changes to harmonize form of address (informal/formal 'you') across materials coming from different cycles,
- requests to harmonize spelling following a spelling reform,

- requests to harmonize decimal and thousand separators across items, and
- changes to improve wording or to correct errors in an item that has not performed well and showed Differential Item Functioning in previous cycles.

The following type of requests were generally rejected:

- preferential changes, improved wording when there are no statistics showing that the item has not performed well in the past,
- punctuation issues,
- capitalization issues (unless outright errors),
- changes that would be against the translation and adaptation guidelines, and
- changes to bring the target version closer to one source version while it already corresponds to the other source version (i.e. changes introducing expressions idiosyncratic to English when a version has been translated from French).

The verifiers' brief for trend verification was to implement the changes that had been requested by the national centre and approved by the referee, if linguistically appropriate – or if not appropriate, suggest a revised wording.

Once changes were verified and implemented in the XLIFF files, verifiers double-checked on the preview that everything appeared correctly in the preview. In principle, no other changes were allowed unless typos or blatant errors were discovered. If the verifier spotted other mistakes that could affect the trend nature of the items, the referee's judgment was called for. At the end of the process, the verifier uploaded the updated XLIFF and Change Request Form files on the portal. The verification step was followed by an internal review by cApStAn. At the end of the process, the files were uploaded on the portal and pushed to Layout Adaptation step.

After the layout adaptation step, the files were pushed to cApStAn for a final check. The main aim of this step was to double-check that all layout issues pointed out during the verification and the review processes had been addressed and to correct any residual issues.

At this stage, the procedure therefore consisted in:

- double-checking that the most important errata (including latest errata released after verification and review) had been implemented
- making sure that the layout issues had been addressed
- addressing any residual issues.

If residual layout issues were found, the relevant files were sent back to ETS for further correction and another check was performed thereafter. At the end of the process, all the files were uploaded on the portal for the national final check and sign-off.

After verification follow-up the countries had a last opportunity to check that all new translations and relevant accepted changes had been implemented correctly and that any residual layout issues, whether raised by the countries themselves or by the verifiers, had been addressed. If errors were still encountered this needed to be commented in the Change Request Form.

The final sign-off from the National Centre ended the trend verification procedure.

## **Questionnaires**

The successful administration of questionnaires in large multinational, multicultural and multilingual surveys depend heavily on their correct adaptation to the national context. The comparability of the data is guaranteed by “asking the same question” in all the Countries/economies and in all the languages, and



to this end, the first task of the PISA Countries/economies was the negotiation of the adaptations, before the translation started.

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 approval 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). Table 7.1 shows a sample questionnaire adaptation spreadsheet from the PISA 2022 Field Trial.

**Table 7.1. Sample of a questionnaire adaptation spreadsheet (QAS) from the PISA 2022 Field Trial**

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
PISA 2022 Questionnaire		PISA 2022 Questionnaire		Questionnaire Adaptation										STEP 3 (PT): TRANSLATION, TEND REVIEW, AND VERIFICATION													
Pre-populated by PISA 2022 Contractor		Pre-populated by PISA 2022 Contractor		STEP 1 (PT): ADAPTATION NEGOTIATION										STEP 2 (PT): TRANSLATION, TEND REVIEW, AND VERIFICATION													
Internal Global Question ID	Internal Question ID	International English version	English translation of your country	French translation of your country	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language
SC000	SC000	International English version	English translation of your country	French translation of your country	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language	Other national language
SC000		Your answer will be kept confidential. They will be verified with answers from other countries to calculate totals and averages in which no school can be identified.		PLEASE NOTE: If you identify high school please use a letter high school rather than answering these questions. If your school is a letter high school answer the question for the 9th grade class.										A NOTE: Si vous établissez est en Lycée, Répondez les questions en vous référant à la classe de 9 <sup>ème</sup> grade. Si vous établissez est en COLLEGE, Répondez les questions en vous référant à la classe de 9 <sup>ème</sup> grade.													
SC000		School identifier code		No further comments.										Agreed													
SC000		School identifier code		Minor linguistic issue										Punctuation issue													
SC000		School identifier code		Corrected ?										Yes													

The verifiers’ brief was to check whether the target questionnaires are linguistically correct and faithful to either the source version (when no adaptation is made) or to 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 accurate,
- 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, and
- to check linguistic correctness (grammar, spelling, etc.) of the entire translated questionnaire.

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

For computer-based questionnaires the verifier applied necessary interventions on OmegaT and documented the rationale for the change in the QAS.

When the verification was completed, the Questionnaire Content Team reviewed the verification feedback and labelled as “requires follow-up” important issues that could potentially affect cross-country comparability. The files were sent back to the country/participant for their review before going through the last passage of Final Check.

The translations of the Global Module Crisis module were produced following a different workflow. cApStAn produced the translated materials through the double translation and reconciliation model. Countries/economies reviewed the translations and requested changes or national adaptations through the QAS. The Questionnaire Content Team assessed the requests and indicated if they were approved or not. The files were then transferred to the verifier who implemented the agreed corrections/updates. There were no special procedures for the verification of the questionnaires adapted from the source versions, from the common base versions or from borrowed versions, 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 Country/economy)

and “quasi-errata” (suggestions for improving the source) was maintained, built up, and used in each successive verification.

As in previous cycles of PISA, 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 four language versions for Spain). Such versions are by necessity entrusted to different verifiers, but when possible, cApStAn’s verification reviewers aimed to review and deliver such versions together, striving to harmonise verification interventions on adaptation issues common to the different language versions.

### ***Adapted versions***

Whenever a country adapted their national version from the English or French source, a common base version, or verified version from the same language borrowed from another country, this was considered an adapted version. The resulting national version was verified using a special procedure for these versions. There were in total 50 CBA adapted versions that were verified using this process.

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

### ***Paper-based test units and booklet shell***

Since no new paper-based units were developed for PISA 2022, PBA Countries/economies that had participated in previous cycles did not have anything new that required translation or adaptation. For these Countries/economies, the units only went through the centralised change management process whereby the Country/economy 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 2022 or that had not participated in one or more of the relevant cycles had to translate or adapt 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 Main Survey Word files using the “track changes” functionality, rather than in OmegaT. The test adaptation spreadsheet was used the same way as in the computer-based verification.

### ***Coding guides***

In PISA 2022, the coding guides were verified separately from the test items, and at a later time. This was necessary since many additions and improvements were made to the master versions after the coder training meetings, long after preliminary versions of the guides had been made available to Countries/economies. As in PISA 2015 and PISA 2018, 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/economies, there was, in addition, one coding guide for New Math, and for those Countries/economies that opted for Financial Literacy and/or Creative Thinking, there were separate coding guides for these domains.

As opposed to the previous cycles, in this cycle the new coding guides were verified using OmegaT. To be able to use the latest version of the translation memories of the cognitive units, the workflows for the coding guides were created only after the cognitive materials were verified. The overall verification procedure was the same as with the cognitive units. The verifiers made corrections as needed in OmegaT, documenting their interventions in the coding guide adaptation spreadsheet (CAS), including selection of

the appropriate intervention category using a drop-down menu. However, there was a significant difference between the verification of the cognitive units and the verification of the coding guides: The translated files for the coding guides were in Main Survey Word format and therefore layout issues had to be corrected manually after the verification process had been completed.

The New Math coding guide went through a full verification in the Field Trial. For the Main Survey, central revisions to reflect updates to the source were made by the Countries/economies in OmegaT together with additional changes which were deemed necessary to correct errors. The verifiers were asked to review both the updates and the edits.

To accommodate the changes to the Creative Thinking coding guides after the Field Trial International Coder Training, the OECD and contractors determined it was important to devote more time to produce updated source versions. Due to time constraints, there was no verification of the Field Trial Creative Thinking coding guides. Instead, a full verification was implemented for the Main Survey.

The Creative Thinking master coding guide was updated after the Field Trial, and the Countries/economies were asked to reflect these updates in their translations. They did this in a newly generated OmegaT project where the translation memories from the revised Main Survey units as well as the translation memories from the Field Trial Creative Thinking coding guides were included. While implementing these central updates in their translations, the Countries/economies also had the opportunity to correct residual errors detected during their review of their Field Trial data.

For Countries/economies that had participated in previous cycles, trend coding guides underwent a similar controlled change request process as for the test units.

### ***Outcomes of the Field Trial verification***

The Test Adaptation Spreadsheets (TAS) and the Questionnaire Adaptation Spreadsheets (QAS) in Excel format were used to document the verification of test units and the questionnaires. For each issue they encountered, verifiers were required to choose from a drop-down list of 14 intervention categories and then explain the details of the issue and of their intervention in a comment.

The predefined intervention categories in the drop-down menus of the TAS and QAS are linked to formulae, which generate statistics on the number and types of verifier interventions in test units, both per language version and per unit. The data is available in detailed form in Appendices 4-8 of this chapter (in Excel format). In this section, some of the data will be presented, together with some figures and graphs.

For reasons of comparability, the data of the translated versions are shown separately from the data of the versions that were adapted from the French or English source versions or from the Chinese or Spanish base version, or from a verified national version of another country. For these adapted versions, the process was different as it was a focused verification of national adaptations proposed by the national centre, rather than a full sentence-by-sentence verification. The results are not comparable with the translated versions where the whole translation was verified sentence by sentence.

The statistics in this section cover national versions of New Mathematics units and Creative Thinking units. The list of language versions is not identical between the two domains for two reasons: some National Centres opted out of the Creative Thinking innovative domain, and for some other countries the Translation Plan was different depending on the domain. Also, some countries opted for a hybrid plan; for example, for the New Mathematics units Bosnia and Herzegovina, Serbia and Montenegro translated a third (one batch) of the units each and adapted the other two thirds, so in the statistics they appear in both tables and graphs.

For each national version included in the analysis, the formulas embedded in each of the TAS produced the following figures:

- the total number of verifier interventions in the 61 New Mathematics units across the 113 language versions;
- the total number of verifier interventions per intervention category in these units; and
- the total number of verifier interventions “requiring follow-up” and related percentage.

In addition, for each unit, data was extracted to obtain:

- the total number of interventions per intervention category (in translated and adapted versions);
- the total number of interventions “requiring follow-up”; and
- the percentage of each type of intervention category vs. the total number of issues reported.

While figures per national version can be informative, they need to be interpreted with care. An illustrative sample of possible scenarios is presented below.

Two versions are of the same generally acceptable quality. One is verified by a strict verifier who extensively comments on even minor errors; another is verified by a more pragmatic verifier who documents only major issues. The statistics might show a great number of interventions in the first version, and considerably less in the other. This difference in verification styles should, however, show in the percentage of interventions “requiring follow-up”, which should be lower than average in the version verified by the “strict” verifier.

One verifier may have reported an “Inconsistency” issue in the TAS every single time the issue appeared. Another verifier may have chosen to report such cases only once, with the note “Corrected throughout the units without further comments” in the verifier comment on the first occurrence. Similarly, one verifier may have reported a recurring issue (e.g. a repeated ‘mistranslation’) each time it occurs, while another verifier might cover that with one generic comment.

Recurring issues, such as missed harmonization of repeated instructions or inconsistency in form of address, generally labelled as “Inconsistency”. If the number of such interventions is very high in a version this may be due to the fact that that trend translations were not considered when translating or adapting the new units.

There may be several separate issues in one sentence/paragraph that the verifier has documented in the same row in the TAS. As only one category can be selected per row, it would be selected according to the most severe issue.

In adapted versions the verifiers are mainly focusing on national adaptations vs. the base and correct implementation of the errata. This explains the fact that these two categories appear to be much higher in adapted versions versus translated versions.

While looking at the total number of interventions does give some indication of the translation quality of the national version, it does not take into account the severity of the issues discovered by the verifier. It makes more sense to look at several combined factors that may serve as indicators for translation quality. One should examine the total number of changes labelled by the Translation Referee as ‘requiring follow-up’ and the number of issues in the more ‘severe’ intervention categories – mistranslation, adaptation issue, matches & patterns, and guideline not followed.

### ***New cognitive items: translated versions***

Even if most of the verifiers rated the translations as very good or good, the verifier interventions were key to maintain the linguistic equivalence to source and correct any residual language issues.

In the translated versions of the New Mathematics units, the categories which revealed the most verification interventions were:

**Minor linguistic issue** – this category is used for typos or other linguistic defect such as spelling, grammar, capitalization, punctuation, etc., that does not significantly affect comprehension or equivalence. Correcting such errors is usually not controversial, and in the Mathematics units 25% (see Figure 7.5) of the verifier’s interventions fall into this category.

**Inconsistency** – typically used for interventions when an element across units (e.g. an instruction or prompt) is inconsistently translated, and it is not intentional or documented as an adaptation. The verifiers’ corrections show 18% in this category, as shown in Figure 7.5.

**Grammar or syntax** – this category was used to document 13% of the verifiers’ interventions (see Figure 7.5). It is used for corrections of grammar mistakes that could affect comprehension or equivalence, e.g. wrong subject-verb agreement, wrong case (inflected languages), wrong verb form, or syntax-related deviation from the source and was used in 13% of the interventions, as shown in Figure 7.5.

The low percentages of corrections of severe translation issues such as mistranslation (6%) or adaptation issues (4%) shows the good quality of the translation (Figure 7.6). No corrections of the matches and patterns were recorded in these units. This deviation from the source of is typically more frequent in Reading literacy units’ literal matches (repetition of the same word or phrase) or a synonymous match (use of a synonym or paraphrase) or patterns in multiple choice items (e.g. all but one option start with the same word, proportional length of responses options) need to be reflected in the target version for valid data measurement and comparison.

**Figure 7.5. Distribution by category of verifier interventions in New Mathematics units (translated versions)**

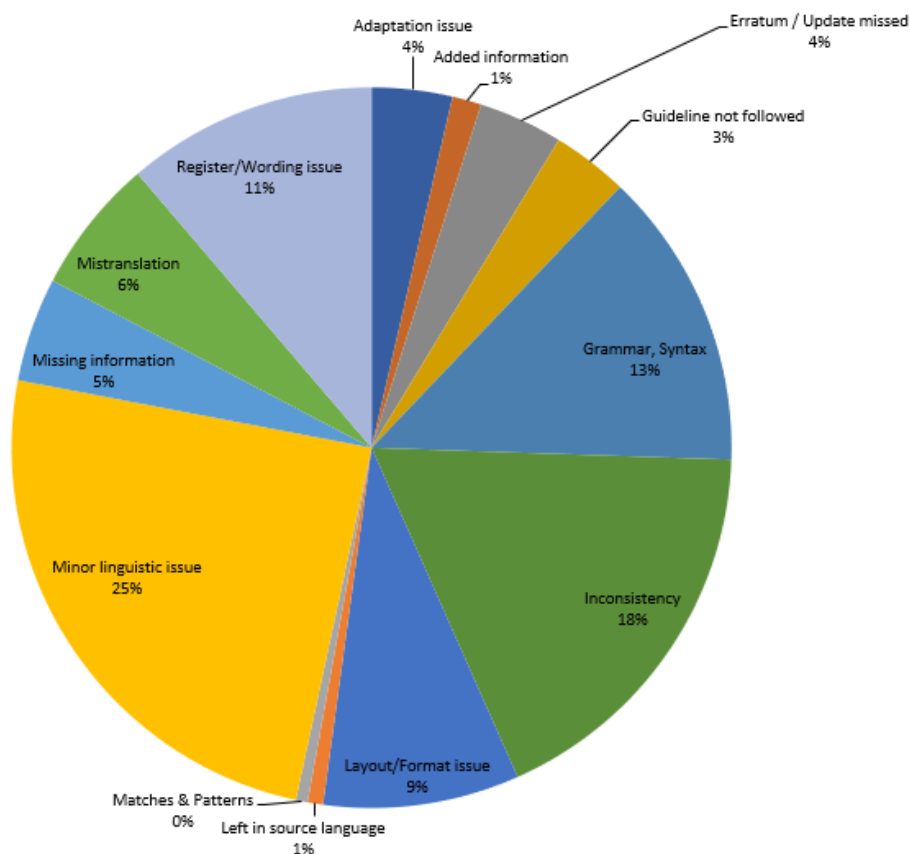
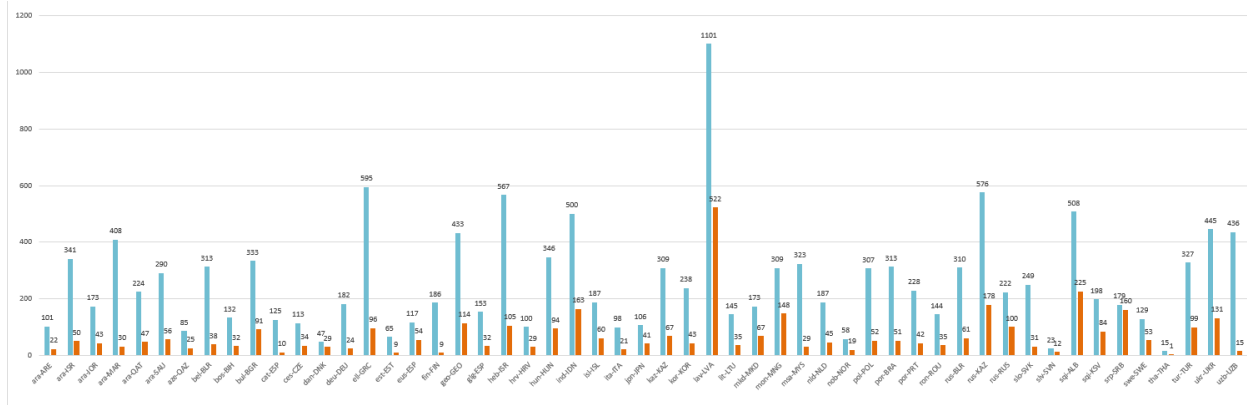


Figure 7.6. Number of issues per national version in New Mathematics units (translated versions)



The outcome of the verification of the Creative Thinking units is similar, with 27% of interventions were for corrections of inconsistent translation and 18% of corrections of minor linguistic issues, as shown in Figure 7.7. The number of issues per national version can be also found in Figure 7.8.

Figure 7.7. Distribution by category of verifier interventions in Creative Thinking units (translated versions)

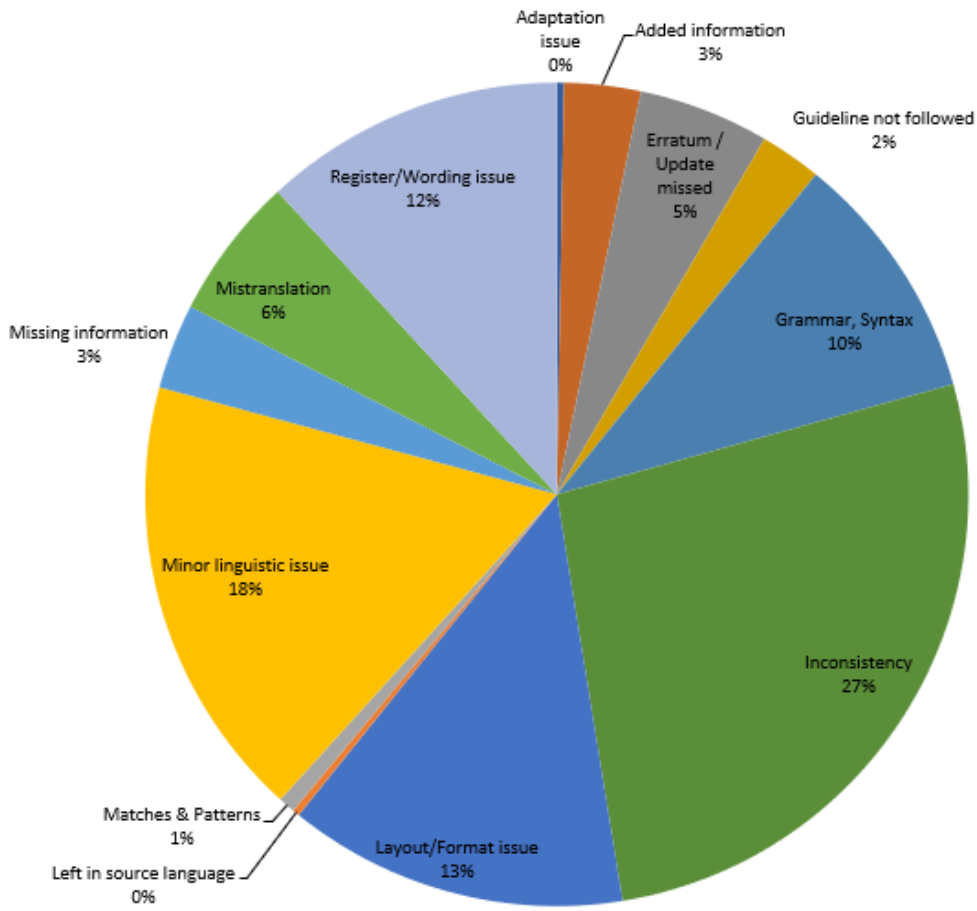
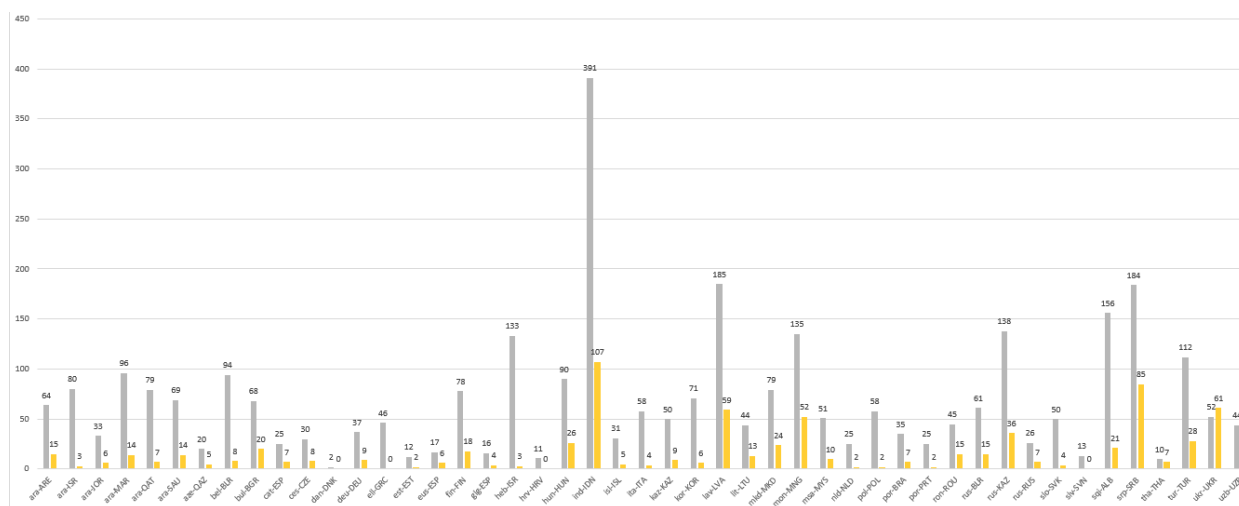


Figure 7.8. Number of issues per national version in Creative Thinking units (translated versions)



### ***New cognitive items: adapted versions***

For the versions adapted from the English or French master version, from the Chinese or Spanish common reference version, or from a borrowed verified national version, the issues identified by verifiers mostly belonged to the following types:

**Adaptation issue** – As shown in Figure 7.9, in 17% of the verifiers' interventions, required adaptation was missed, materials were not adapted at all or poorly adapted; adaptations was not correctly or consistently implemented. For example, the adaptation documented in the TAS was not implemented as described in the XLIFF file, or implemented only in some occurrences; adaptation or change proposed by national centre was not acceptable (e.g. it added information not present in the source or made the national version easier or more difficult). Typical examples of adaptation issues in adapted versions are: missed adaptation of spelling and typographic conventions (e.g. UK to US English spelling, date formatting, decimal and thousands separators), fictitious character names not adapted to local context, etc.

**Inconsistency** – similar to the translated version, 19% of the corrections fall into this category (See Figure 7.9).

Minor linguistic issues were corrected in 14% of the interventions, errata were corrected in 12% of the interventions and layout or formatting such as emphasis (bold, italics, underline) was adjusted in 10% of the interventions in the adapted versions of the New Mathematics units, as shown in Figure 7.9.

Figure 7.9. Distribution by category of verifier interventions in New Mathematics units (adapted versions)

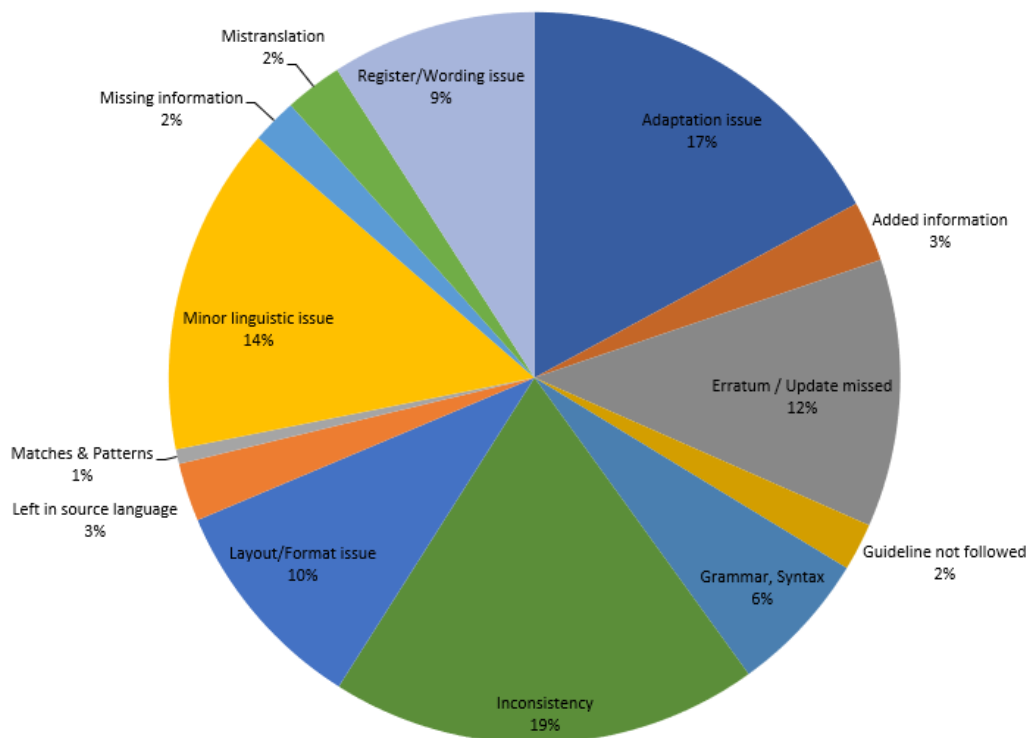
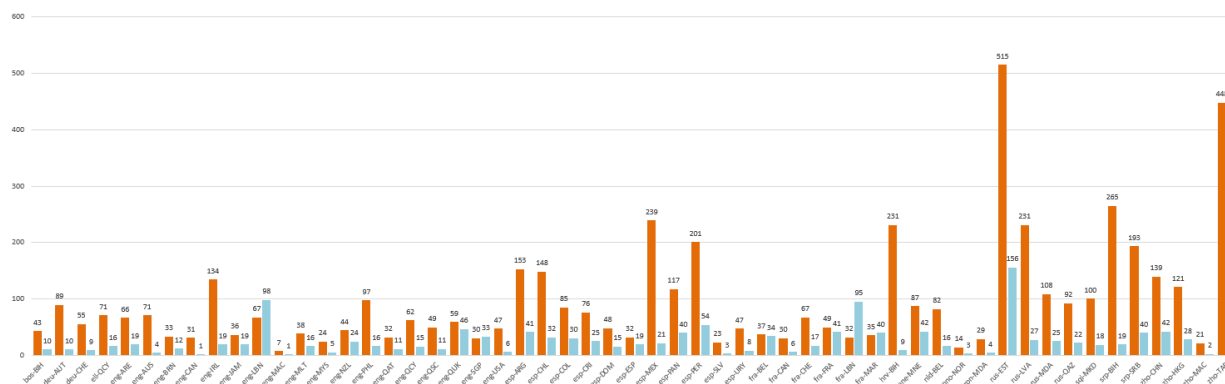


Figure 7.10. Number of issues per national version in New Mathematics units (adapted versions)



For Creative Thinking, inconsistencies were harmonized in 21% of the verifiers’ interventions, errata were corrected by the verifiers in 18% of their interventions, and register, wording and minor linguistic issues were corrected in 12% of the recorded interventions, as per Figure 7.11. In addition, Figure 7.12 presents a breakdown of issues per national version.



Figure 7.11. Distribution by category of verifier interventions in Creative Thinking units (adapted versions)

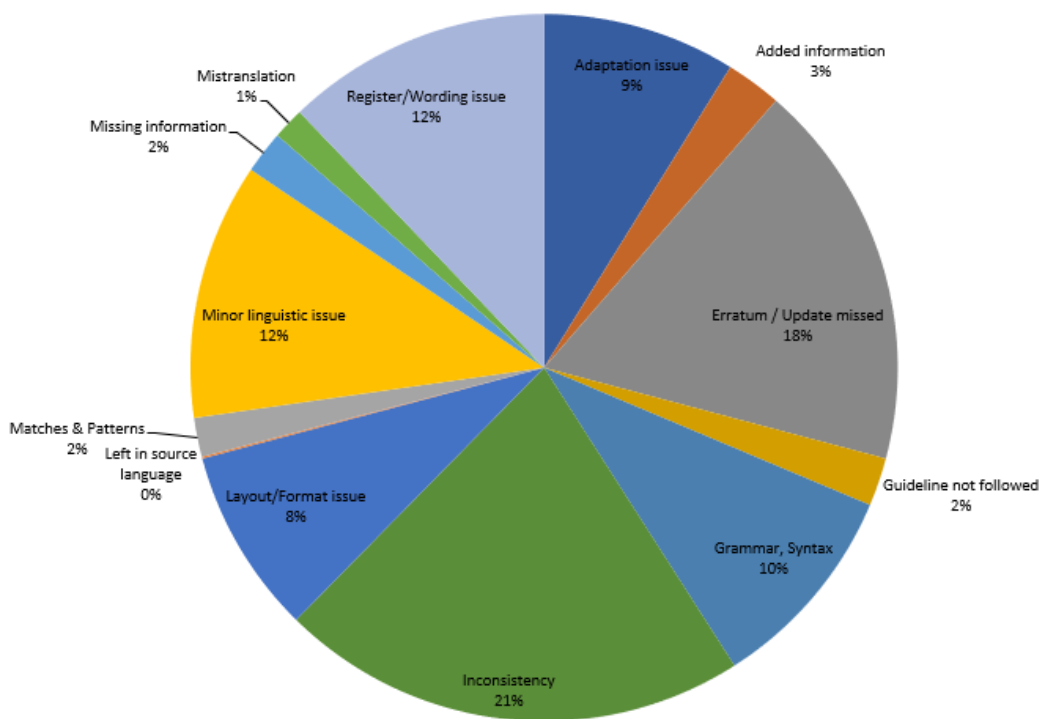
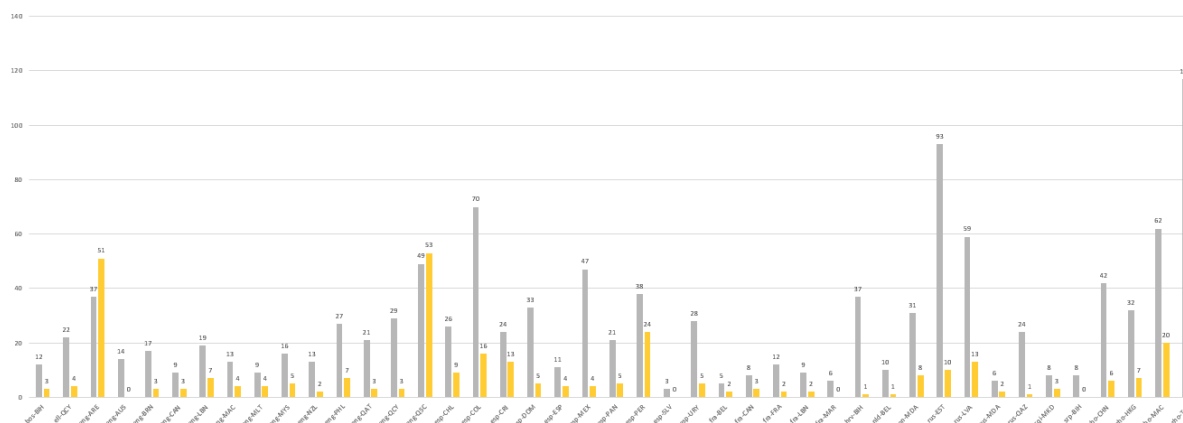


Figure 7.12. Number of issues per national version in Creative Thinking units (adapted versions)



## Main survey verification

### Cognitive units

As in PISA 2018, no major changes were made in the master versions after the Field Trial (apart from entire units or items being dropped) in PISA 2022. The changes that Countries/economies requested to their Field Trial instruments, for example based on poor performance or differential item functioning in the Field Trial, or the detection of residual “outright errors” needed to be verified and centrally implemented together with the implementation of the FT-to-MS errata. These errata included errata discovered after the

last release of the Field Trial errata document and central Field Trial to Main Survey updates. This process was similar to the centralised change management used to control changes in trend: Countries/economies requested changes, and the verifiers implemented centrally those changes that were approved by the translation referee. The Countries/economies did not have editing access to their units or questionnaires at this stage.

The trend items and new items followed the same workflow although the process for New and Trend materials was slightly different.

The Main Survey preparation started after the Main Survey item selection was confirmed. The release of the Main Survey workflows was linked to the data release, so the timeline strongly depended on the compliance of the National Centres in submitting their Field Trial data.

During the Main Survey review countries were asked to carefully review any items that did not perform well in the Field Trial and try to identify whether this country-item interaction was language-driven. Such items were highlighted in red in the "item feedback form" (IFF) in Excel format and indicated by a "YES" in the "Flagged Item?" column.

In case the National Centre spotted residual errors, they had the opportunity to request changes to the translation. Changes had to be requested in the IFF where countries were asked to enter: a short description of the error, the location of the error (e.g. segment number), the English or French source for that segment, the original Field Trial wording, a back-translation of the original Field Trial wording and the proposed corrected Main Survey wording.

There was one item feedback form for all cognitive items with a separate tab for each domain (New Maths, Trend Reading, Trend Science, Trend Math, XYZ and Trend-New Financial Literacy). The IFF also included 2 Instructions tabs describing in detail the process for the new and trend materials. At verification 2 additional columns indicating the dropped items and the Main Survey errata were added.

There was one single workflow for all the Core instruments for the Field Trial to Main Survey verification which included two Referee review steps, one *before* the verification which was used to review national centre requests for changes in the trend materials, and one *after* verification, to flag any major issue, as usual.

All national centre requests were reviewed by the verifier, who double-checked (i) whether it was an outright error or a preferential change, and (ii) whether the proposed Main Survey wording was still equivalent to the source and linguistically correct. As a general principle, for the trend materials the principle of identicalness of trends was applied and any preferential change or change to a non-flagged item was generally not agreed by the Referee and therefore not implemented by the verifier. Agreed corrections were corrected in the XLIFF files by the verifier. Additionally, verifiers were responsible for implementing all Field Trial to Main Survey errata, that is errata which were discovered between the Field Trial and the Main Survey.

Countries did not have access to the XLIFF files at any point of the process; all changes were implemented centrally by cApStAn verifiers. Countries could nevertheless consult the unit previews and DIF reports at different stages of the process, to make sure their requested changes and the Field Trial to Main Survey errata were correctly implemented during verification.

The general guideline of correcting only outright errors (and, more generally, the concept of "outright error") was not understood and accepted the same way by all countries. Some only requested a few justified changes, others called for a more extensive revision of the units (e.g. Kazakhstan, Mongolia).

## **Questionnaires**

As in Cognitive items, no content-related changes were made in Questionnaires items that made it to the Main Survey. The changes in the questionnaires before the Main Survey were mainly structural. Full

questions and response options within items were omitted, the order of questions was changed. Finally, a couple of updates in two questions and an introductory part were implemented centrally by ETS.

The structural changes were implemented by ETS in the QAT for the countries that administered the questionnaires on computer. For the paper-based questionnaires, the National Centres reflected these changes in their materials in Word format before generating the final questionnaires in PDF format.

The procedure was similar to the procedure for the cognitive units. The few content changes such as the addition of the consistency checks for scale questions were considered as errata and were added to the necessary update in the year of administration in SC002 and the Field Trial to Main Survey errata. A tab for the documentation of these updates, the Main Survey Questionnaire Change Request Form was added in the QAS (Table 7.2). The QAS also contained the locked Field Trial QAS tabs for reference, well as a tab with an example of correctly documented Main Survey Changes.

In the Main Survey Questionnaire Request Form countries could also request other updates due to objective major modifications (e.g. changes in the school programs on national level), or ask for correction of errors in items showing strange behaviour in the Field Trial data. They were advised against any changes in items that worked well in the Field Trial.

The Questionnaire Team at ETS reviewed the documented updates and possible requests for corrections of errors and recommended their implementation when applicable.

At verification stage, the verifier checked the linguistic correctness of the update in the target language and implemented centrally to the questionnaires in XLIFF the agreed changes. In the step after this implementation, countries could review it in the QAT, and reported in the QAS if any residual issues needed to be addressed.

The same procedure was followed for the PBA materials, with the difference that the National Centres reflected the recommended updates and agreed corrections in the questionnaires administered on paper.

**Table 7.2. Main Survey Questionnaire Change Request Form in the QAS**

PISA 2022: MS Questionnaires Change Request Form							Step 6.2 [MS] – Approve Changes		
Step 6.1 [MS] – Identify Changes							ETS Questionnaire Content		
PISA Centre				ETS Translation			ETS Questionnaire Content		
6.1: Fill in these 5 columns for all corrections				6.1: Fill in these 3 columns for all translation corrections			6.2: Comments on correction requests		
Questionnaire	Question ID	Item ID	Type of Correction	Description of correction	ET Translation, in National Language (Full segment)	Requested Translation for MS, in National Language (Full segment)	Back-translation of requested MS Translation, in ENGLISH	6.2 Comments on correction requests	6.2 Approval Status
(Select from the dropdown)	(e.g., ST001)	(e.g., ST001QA01TA01)	(Select from the dropdown menu)						
TCQ	TC258	TC258E01	ERRATUM	Addition of Consistency Check for Scale Questions to avoid items being listed as unanswered in the Questionnaire. PISA Centres should translate the text in Column H into their national language, and include the updated translation in	N/A	"0"이라고 응답하기 위해서는 눈금의 슬라이더(조절 단추)를 "0"의 위치로 이동하세요.	To enter a response of "0" (zero) for a question, please move the slider to the "0" position on the scale.		AGREED
TCQ	TC259	TC259E01	ERRATUM	Addition of Consistency Check for Scale Questions to avoid items being listed as unanswered in the Questionnaire. PISA Centres should translate the text in Column H into their national language, and include the updated translation in	N/A	"0"이라고 응답하기 위해서는 눈금의 슬라이더(조절 단추)를 "0"의 위치로 이동하세요.	To enter a response of "0" (zero) for a question, please move the slider to the "0" position on the scale.		AGREED
TCQ	TC261	Description	DELETION	Description removed from this item. Please review item screen to ensure that the description has been properly deleted.	Please consider your employment status at this school and for all of your teaching employments together.	Deleted	N/A		AGREED
TCQ	TC261	Instruction	ERRATUM	Incorrect instruction text. Instruction changed to (Please select one response.) for this item. PISA Centres should translate the text in Column H into their national	(Please select one response in each row.) (각 항목에서 하나를 선택하십시오.)	(하나를 선택하십시오.)	(Please select one response.)		

### Coding Guides

The coding guides for the new cognitive items were translated and verified in XLIFF format, therefore the Main Survey updates and corrections of the errata followed the same procedure as the instruments.

For the Main Survey, the countries were asked to produce Main Survey versions of their trend coding guides starting from their final Field Trial versions, reflecting all applicable revisions made in the master

versions. Separate Main Survey master versions were produced for PB and CB countries. The Field Trial to Main Survey revisions that countries were asked to reflect were of the following types:

- Removing scoring sections of items that did not make it to the MS
- Making edits in the cover, footers and introduction
- Reflecting Field Trial to Main Survey revisions the test developers made in the scoring sections, e.g. modifications in the scoring instructions or addition/removal of sample responses

Master versions with tracked changes were released to countries and they were asked to reflect all the Field Trial to Main Survey revisions in their national version (using track changes) before submitting them for verification.

The verification of the New Mathematics coding guides was a focused verification only on revisions and concerned all CBA countries which had previously translated/adapted the Field Trial guides. Similar to the Main Survey verification of the cognitive units, countries could request changes either to correct residual errors or, in some cases, to modify the scoring instructions based on coder feedback or because the item showed differential functioning in the Field Trial, and a potential reason for this had been identified in the scoring instructions. If the National Centre did not request changes in the trend guides, these were not verified at all and the few revisions from Field Trial to Main Survey in Trend were left under National Centre responsibility.

The Main Survey verification procedure of coding guides was similar to that of cognitive items and followed the same workflow on the portal: countries could request justified changes to trend in the "Coding guide feedback form" in Excel format (CFF). The main difference compared to cognitive units was that all changes were implemented by the countries, while for cognitive units the countries did not have access to the files at any point, and verifiers made the changes in their New and Trend guides.

The translation memories from their final cognitive instruments were included in the national OmegaT packages, thus the quotations from the test items were identical with the instruments. The translation memories from their Field Trial coding guides were also included, and for the source segments that stayed identical as in the Field Trial, the translation was auto populated. The target segments for which the source segments changed in Main Survey were empty, while the translation from the Field Trial was available in the fuzzy matches pane. The country could update the Main Survey coding guides and correct the errata using the existing translation, as well as the consistency tools in the OmegaT. For the adapted versions, Chinese and Spanish Main Survey common reference versions were produced, and their translation memories from the Main Survey instruments and Field Trial coding were included in their national packages. These countries had to make sure that their adaptations were correctly reflected in the updated segments.

The completed forms and revised XLIFF and Word files were then submitted to Translation Referee for approval. Once the Referee had finished the review of the CFF, the files moved to verification. For the New Mathematics coding guide, the Referee review took place after verification. The verification and Referee review outcomes were documented in the same CFF. At verification, the DIF report was checked to make sure no undocumented changes were made.

When the National Centre did not request any changes to trend, a spot check was performed to their coding guides. If such changes were discovered the National Centre was asked either to provide a complete documentation, or to start over the preparation of the Main Survey guides (for example, if by mistake an outdated version was used as starting point).

For the countries that decided to use the master version as such either in ENG or FRA (e.g. Germany), the guides were not verified.

## **Errata management during Main Survey**

### *Errata in Cognitive materials*

Before the Main Survey preparation started, an all-in-one Field Trial to Main Survey Errata Document was released to the countries. This document included the errata released after the Field Trial and during the Main Survey review process for the Cognitive units. Countries/economies did not need to request the implementation of Main Survey errata. All of these errata were systematically checked and corrected by the verifiers, at verification step. At Post-Verification review step, the Countries/economies had to make sure that all released Main Survey errata have been addressed in a satisfactory way. If any Main Survey erratum was missed at verification, Countries/economies needed to indicate this in the CFF Coding Guides Follow-up Form, providing 1) the errata reference (from col. "Reference") and 2) the corrected version that the verifier should implement (whole segment). It was then addressed at Final Check.

The errata list included separate lists of errata identifying the errors in the English and the French source, as well as a separate tab with one erratum to be corrected in the source version in French in trend Reading item R549Q12: the wrong option was deleted in Source after selection for PISA 2018. This did not apply to National French versions for countries who participated to PISA 2018. The NCs were instructed to refer to that document to double-check if any of the errors listed in that file affected their national version if the reconciler had relied on the translation produced from French for a particular unit or section.

### *Errata in Questionnaires*

The errata that were identified and approved for correction by the contractors before the Main Survey were documented in the Questionnaire Change Request Form in the Main Survey QAS, and the Countries/economies provided the corrected version in it. The verifiers then implemented the correction at verification step. The Countries/economies checked that the implementation was correct and documented residual issues, addressed by verifier at final check.

## **Suggestions for the future**

The suggestions and lessons learnt in the PISA 2018 were taken on board and the process was significantly improved in PISA 2022. The major break-through in PISA 2022 was the use of OmegaT for translation, adaptation and verification of the PISA instruments. The PISA 2022 portal presented a clearer overview, a straightforward layout and yet a number of improved functionalities over the previous cycle. The coding guides for the new cognitive units were translated and verified in XLIFF format in OmegaT, benefitting from the translation memories from the verified cognitive units. The questionnaires were adapted in QAT, national master in XLIFF was exported from the QAT and translated in OmegaT. The Main Survey procedures for the cognitive and questionnaires got closer – a Questionnaire Change Request Form was used in the Main Survey verification.

At the conclusion of this process, we have the following specific recommendations in three areas.

### *Communication with countries and processes*

In this cycle, communication with countries worked well. The trainings and webinars, the video tutorials, the User Guides, the questions, and answers section on the portal all contributed to clarify the different tasks to be performed at country level. In addition, at the end of each step, the NPM received an email with the instructions for the next step. On the other hand, not all national centres consulted and followed the instructions as expected. This could be due to various factors, such as (i) national centres not finding the instructions, (ii) national centre delegating the task to a person without forwarding the instructions (iii) user not understanding the instructions. The complexity of the PISA procedures and workflows may be rendered more understandable to the users if they are explained in pre-recorded webinar sessions that the

Countries/economies should watch before the live sessions during the face-to-face trainings and/or live webinar sessions. The trainings and the live sessions would then focus on Countries/economies' questions, issues, hands-on exercises, and particular difficulties.

### *File management*

Although the PISA 2022 CBA Countries/economies could benefit from powerful translation memory management of the open-source CAT tool OmegaT in PISA 2022, version management issues were still a challenge in this cycle, i.e. national centre uploading an outdated version back to the workflow and pushing it forward, or national centre editing an outdated version and pushing it forward, losing the feedback provided in a previous step. A team OmegaT project may resolve this issue, where the online OmegaT package is automatically opened at each step of the workflow.

### *Errata management*

Although in this cycle the errata management process was improved over PISA 2018, it was still observed that corrections were not implemented in the materials by the national centres. In the next cycle the errata management could also benefit from the use of OmegaT team project approach: at each source update, the target segments would appear untranslated, and the existing (outdated) translation from the translation memory would be shown in the fuzzy matches for reference. The user would then need to correct the translation so that it matches the updated source version.

**Table 7.3. Overview of Testing and Questionnaire Items**

Table/Figure	Title
Table 7.1	Sample of a questionnaire adaptation spreadsheet (QAS) from the PISA 2022 Field Trial
Table 7.2	Main Survey Questionnaire Change Request Form in the QAS
Figure 7.3	Sample of a test adaptation spreadsheet (TAS) from the PISA 2022 Field Trial
Figure 7.6	Distribution by category of verifier interventions in New Mathematics units (translated versions)
Figure 7.7	Number of issues per national version in New Mathematics units (translated versions)
Figure 7.8	Distribution by category of verifier interventions in Creative Thinking units (translated versions)
Figure 7.9	Number of issues per national version in Creative Thinking units (translated versions)
Figure 7.10	Distribution by category of verifier interventions in New Mathematics units (adapted versions)
Figure 7.11	Number of issues per national version in New Mathematics units (adapted versions)
Figure 7.12	Distribution by category of verifier interventions in Creative Thinking units (adapted versions)
Figure 7.13	Number of issues per national version in Creative Thinking units (adapted versions)

StatLink  <https://stat.link/hxtfy5>

## Annex 7.A. Translation items

**Annex Table 7.A.1. Chapter 7: Translation and Verification of PISA 2022 Survey Materials**

Tables	Title
Table 7.A.2	Verified language versions of the PISA 2022 materials
Table 7.A.3	List of New Mathematics units in PISA 2022 Field Trial
Table 7.A.4	List of Field Trial New Mathematics not administered in the Main Survey
Table 7.A.5	List of Creative Thinking units in PISA 2022 Field Trial
Table 7.A.6	Translation procedures reported by national centres in the translation plan

**Annex Table 7.A.2. Verified language versions of the PISA 2022 materials**

PISA Participant	Language	Code	Last Cycle	Mode	PBA->CBA	Adpt	CT	FL	ICQ	TCQ	WBQ	UH	PAQ
Albania	Albanian	sqi-ALB	2018	CBA			Y**		Y	Y			
Argentina	Spanish	esp-ARG	2018	CBA	Y	Y			Y				
Australia	English	eng-AUS	2018	CBA		Y	Y		Y	Y			
Austria	German	deu-AUT	2018	CBA		Y			Y			Y	
Azerbaijan (Baku city only)	Azerbaijani	aze-QAZ	2018	CBA			Y			Y			
Azerbaijan (Baku city only)	Russian	rus-QAZ	2018	CBA		Y	Y			Y			
Belgium	French	fra-BEL	2018	CBA		Y	Y		Y			Y	
Belgium	Dutch	nld-BEL	2018	CBA		Y	Y		Y				Y
Bosnia and Herzegovina	Bosnian	bos-BIH	2018	CBA		Y	Y					Y	
Bosnia and Herzegovina	Croatian	hrv-BIH	2018	CBA		Y	Y					Y	
Bosnia and Herzegovina	Serbian	srp-BIH	2018	CBA		Y	Y					Y	
Brazil	Portuguese	por-BRA	2018	CBA			Y	Y	Y	Y	Y		Y
Brunei Darussalam	English	eng-BRN	2018	CBA		Y	Y		Y				
Bulgaria	Bulgarian	bul-BGR	2018	CBA			Y	Y	Y				
Cambodia	Khmer	khm-KHM	PISA-D	PBA									
Canada	English	eng-CAN	2018	CBA		Y	Y	Y					
Canada	French	fra-CAN	2018	CBA		Y	Y	Y					
Chile	Spanish	esp-CHL	2018	CBA		Y	Y	Y	Y	Y	Y	Y	Y
B-S-J-Z (China)	Chinese (simpl.)	zho-CHN	2018	CBA		Y	Y		Y	Y			
Colombia	Spanish	esp-COL	2018	CBA		Y	Y			Y			Y
Costa Rica	Spanish	esp-CRI	2018	CBA		Y	Y	Y	Y	Y	Y	Y	Y
Croatia	Croatian	hrv-HRV	2018	CBA			Y		Y				Y
Cyprus	Greek	ell-QCY	2018	CBA		Y	Y						
Cyprus	English	eng-QCY	2018	CBA		Y	Y						

PISA Participant	Language	Code	Last Cycle	Mode	PBA->CBA	Adpt	CT	FL	ICQ	TCQ	WBQ	UH	PAQ
Czech Rep.	Czech	ces-CZE	2018	CBA			Y	Y	Y			Y	
Denmark	Danish	dan-DNK	2018	CBA			Y	Y	Y			Y	
Dominican Republic	Spanish	esp-DOM	2018	CBA		Y	Y		Y	Y			Y
El Salvador	Spanish	esp-SLV	NEW	CBA		Y	Y						
Estonia	Estonian	est-EST	2018	CBA			Y		Y				
Estonia	Russian	rus-EST	2018	CBA		Y	Y		Y				
Finland	Finnish	fin-FIN	2018	CBA			Y		Y				
France	French	fra-FRA	2018	CBA		Y	Y				Y		
Georgia	Georgian	geo-GEO	2018	CBA					Y	Y			Y
Germany	German	deu-DEU	2018	CBA			Y		Y	Y		Y	Y
Greece	Greek	ell-GRC	2018	CBA			Y		Y				
Guatemala	Spanish	esp-GTM	PISA-D	PBA		Y							
Hong Kong (China)	Chinese (trad.)	zho-HKG	2018	CBA		Y	Y		Y	Y	Y		Y
Hungary	Hungarian	hun-HUN	2018	CBA			Y	Y	Y		Y		
Iceland	Icelandic	isl-ISL	2018	CBA			Y		Y			Y	
India (Chandigarh)	English	eng-QIN	NEW	PBA		Y							
India	Hindi	hin-QIN	NEW	PBA									
Indonesia	Bahasa Indonesia	ind-IDN	2018	CBA			Y	Y					
Ireland	English	eng-IRL	2018	CBA		Y			Y		Y		Y
Israel	Arabic	ara-ISR	2018	CBA			Y		Y				
Israel	Hebrew	heb-ISR	2018	CBA			Y		Y				
Italy	Italian	ita-ITA	2018	CBA			Y	Y	Y				Y
Jamaica	English	eng-JAM	NEW	CBA		Y	Y						Y
Japan	Japanese	jpn-JPN	2018	CBA					Y				
Jordan	Arabic	ara-JOR	2018	CBA	Y		Y		Y				
Kazakhstan	Kazakh	kaz-KAZ	2018	CBA			Y		Y				
Kazakhstan	Russian	rus-KAZ	2018	CBA		Y	Y		Y				
Korea	Korean	kor-KOR	2018	CBA			Y		Y	Y			Y
Kosovo	Albanian	sqi-KSV	2018	CBA		Y				Y			
Latvia	Latvian	lav-LVA	2018	CBA			Y		Y				Y
Latvia	Russian	rus-LVA	2018	CBA		Y	Y		Y				Y
Lebanon	English	eng-LBN	2018	CBA	Y	Y	Y		Y	Y			
Lebanon	French	fra-LBN	2018	CBA	Y	Y	Y		Y	Y			
Lithuania	Lithuanian	lit-LTU	2018	CBA			Y		Y				
Macao (China)	English	eng-MAC	2018	CBA		Y	Y		Y	Y	Y		Y
Macao (China)	Chinese (trad.)	zho-MAC	2018	CBA		Y	Y		Y	Y	Y		Y
Malaysia	Malaysian	msa-MYS	2018	CBA			Y	Y	Y	Y			
Malaysia	English	eng-MYS	2018	CBA		Y	Y	Y	Y	Y			
Malta	English	eng-MLT	2018	CBA		Y	Y		Y				
Malta	Maltese	mlt-MLT	2018	CBA			Y		Y				



PISA Participant	Language	Code	Last Cycle	Mode	PBA->CBA	Adpt	CT	FL	ICQ	TCQ	WBQ	UH	PAQ
Mexico	Spanish	esp-MEX	2018	CBA		Y	Y			Y	Y		
Moldova	Romanian	ron-MDA	2018	CBA	Y	Y	Y						
Moldova	Russian	rus-MDA	2018	CBA	Y	Y	Y						
Mongolia	Mongolian	mon-MNG	NEW	CBA			Y						
Montenegro	Montenegrin	mne-MNE	2018	CBA		Y							
Morocco	Arabic	ara-MAR	2018	CBA			Y		Y	Y			
Morocco	French	fra-MAR	2018	CBA		Y	Y						
Netherlands	Dutch	nld-NLD	2018	CBA			Y	Y			Y	Y	
New Zealand	English	eng-NZL	2018	CBA		Y	Y		Y		Y		
North Macedonia	Albanian	sqj-MKD	2018	CBA	Y	Y	Y						
North Macedonia	Macedonian	mkd-MKD	2018	CBA	Y		Y						
Norway	Bokmål	nob-NOR	2018	CBA				Y				Y	
Norway	Nynorsk	nno-NOR	2018	CBA		Y		Y				Y	
Panama	Spanish	esp-PAN	2018	CBA		Y	Y		Y	Y	Y		Y
Paraguay	Spanish	esp-PRY	PISA-D	PBA		Y							
Peru	Spanish	esp-PER	2018	CBA		Y	Y	Y		Y			
Philippines	English	eng-PHL	2018	CBA		Y	Y						
Poland	Polish	pol-POL	2018	CBA			Y	Y	Y				
Portugal	Portuguese	por-PRT	2018	CBA			Y	Y		Y			Y
Qatar	Arabic	ara-QAT	2018	CBA			Y						
Qatar	English	eng-QAT	2018	CBA		Y	Y						
Romania	Romanian	ron-ROU	2018	CBA	Y		Y		Y				
Saudi Arabia	Arabic	sau-ARA	2018	CBA	Y		Y						Y
Serbia	Serb (Ekavian)	srp-SRB	2018	CBA			Y	Y					
Singapore	English	eng-SGP	2018	CBA		Y	Y		Y				
Slovak Rep.	Slovak	slo-SVK	2018	CBA			Y		Y			Y	
Slovenia	Slovenian	slv-SVN	2018	CBA			Y		Y		Y	Y	
Spain	Basque	eus-ESP	2018	CBA			Y	Y	Y		Y		
Spain	Galician	glg-ESP	2018	CBA			Y	Y	Y		Y		
Spain	Castilian	esp-ESP	2018	CBA		Y	Y	Y	Y		Y		
Spain	Catalan	cat-ESP	2018	CBA			Y	Y	Y		Y		
Sweden	Swedish	swe-SWE	2018	CBA					Y				
Switzerland	French	fra-CHE	2018	CBA		Y			Y				
Switzerland	German	deu-CHE	2018	CBA		Y			Y				
Chinese Taipei	Chinese (trad.)	zho-TAP	2018	CBA		Y	Y		Y				
Thailand	Thai	tha-THA	2018	CBA			Y		Y				
Turkey	Turkish	tur-TUR	2018	CBA			Y		Y				
Ukraine	Ukrainian	ukr-UKR	2018	CBA	Y		Y		Y			Y	
UAE	Arabic	ara-ARE	2018	CBA			Y	Y		Y	Y		
UAE	English	eng-ARE	2018	CBA		Y	Y	Y		Y	Y		

PISA Participant	Language	Code	Last Cycle	Mode	PBA->CBA	Adpt	CT	FL	ICQ	TCQ	WBQ	UH	PAQ
United Kingdom (excl. Scotland)	English	eng-QUK	2018	CBA		Y			Y				
United Kingdom (Scotland)	English	eng-QSC	2018	CBA		Y	Y		Y				
United States	English	eng-USA	2018	CBA		Y		Y	Y	Y		Y	
Uruguay	Spanish	esp-URY	2018	CBA		Y	Y		Y				
Uzbekistan	Uzbek	uzb-UZB	NEW	CBA			Y						
Viet Nam	Vietnamese	vie-VNM	2018	PBA									

Note:

This list reflects countries and economies that submitted instruments for verification. For actual participation status, please refer to Table 1.1 in this report.

Y" stands for "Yes" in this table.

### Annex Table 7.A.3. List of New Mathematics units in PISA 2022 Field Trial

Batch	Unite identifier	Unit	
Batch 1	MA101	Building Blocks	
	MA102	Buying a Wardrobe	
	MA103	Calculation Program	
	MA104	Car Purchase	
	MA105	Clean Energy	
	MA106	DVD Sales	
	MA107	Field OF Vision	
	MA108	Fountains	
	MA109	Headache Medicine	
	MA112	Metabolism	
	MA125	Painting A Room	
	MA128	Salinity OF Water	
	MA153	Gears	
	MA159	Spinners	
	MA160	University Student Employment	
	MA161	Forested Areas	
	MA162	Urban Population	
	Batch 2	MA110	Headphone Order
		MA111	Health App
MA113		Heart Rate	
MA114		Honey	
MA115		Iceberg	
MA116		International School	
MA117		Mixing Paint	
MA118		Moving Truck	
MA119		Music Survey	
MA120	Number Cubes		

Batch	Unite identifier	Unit
	MA121	Mobile Phone Reviews
	MA122	Pool Cover
	MA123	Solar System
	MA124	Zedland Topography
	MA126	Robot
	MA127	Predicting Height
	MA129	Shelving Unit
	MA130	Sleep and Reaction Time
	MA131	Travelling by Train
	MA132	Water Temperature
Batch 3	MA133	Arranging Tables
	MA134	Car and Bicycle Ownership
	MA135	Electric Bicycle
	MA136	Movie Rewards
	MA137	Football Tournament
	MA138	Shoe Sizes
	MA139	Tablet Cover
	MA140	Walk to School
	MA141	Water Bill
	MA142	Water Reservoir
	MA143	Wild Bird Food
	MA144	Yogurt
	MA145	Shadows
	MA146	Fuel
	MA147	Aeroplane Tickets
	MA148	Chance of Rain
	MA149	Floor Area
	MA150	Triangular Pattern
	MA151	Moving Out
	MA152	The Better Deal
	MA154	Company Logo
	MA156	Points
	MA157	Tyres
	MA158	Eye Colour
Batch 6A	M905	Tennis Balls
	M919	Fan Merchandise
	M943	Arches
	M953	Flu Test
	M954	Medicine Doses
Batch 6B	M936	Seats in a Theatre
	M939	Racing
	M948	Part Time Work
	M961	Chocolate

Batch	Unit identifier	Unit
	M967	Wooden Train Set

#### Annex Table 7.A.4. List of Field Trial New Mathematics not administered in the Main Survey

New Mathematics	Dropped item/unit in MS
MA101 Q03	Dropped item
MA103 Q03	Dropped item
MA104	Dropped unit
MA106	Dropped unit
MA114 Q02	Dropped item
MA117 Q05	Dropped item
MA118	Dropped unit
MA122	Dropped unit
MA123 Q03	Dropped item
MA126 Q01	Dropped item
MA136 Q01	Dropped item
MA137 Q02	Dropped item
MA144 Q02	Dropped item
MA156 Q02	Dropped item
MA159	Dropped unit

#### Annex Table 7.A.5. List of Creative Thinking units in PISA 2022 Field Trial

Unit Identifier	Unit
T200	Science Fair Poster
T240	Space Comic
T300	Illustration Titles
T350	Book Covers
T360	Moving Backward
T370	2983
T400	Save the Bees
T420	Clean Oceans
T450	Music Festival
T500	Wheelchair Accessible Library
T520	Painting Class
T540	Infographics
T550	Experiment Kit
T560	The Ball
T570	Robot Story
T610	Food Waste
T620	Paper Products
T630	Carpooling
T680	Rubber Ducks Game

Unit Identifier	Unit
T690	Save the River
T700	The Exhibit

**Annex Table 7.A.6. Translation procedures reported by national centres in the translation plan**

Type	Cognitive Items	Questionnaires
Double translation from English and French source versions	17	18
Double translation from English source version with cross-checks against the FRA source version	8	8
Double translation from English source version only	30	39
Adaptation from one of the source versions	25	25
Adaptation from a borrowed verified version or from a common base version	29	23
Double translation from English source version with cross-checks against the Spanish common reference version	3	2

## Notes

1. A translation memory is a database that stores sentences, paragraphs or segments of text that have been translated before.
2. Following Note 4.1 to the PISA 2022 Technical Standards.

## Annex 7.B. Verifier interventions

Annex Table 7.B.1. Chapter 7: Verifier intervention categories

Category	Description
OK	No intervention is needed. The verifier has checked and confirms that the text element or segment is equivalent to source, linguistically correct, and – if applicable – that it conforms to an explicit translation/adaptation guideline. This category may also be used to report an appropriate but undocumented adaptation.
Added information	An information is present in the target version but not in the source version, e.g. an explanation between brackets of a preceding word.
Missing information	An information is present in the source version but omitted in the target version.
Matches and patterns	A literal match (repetition of the same word or phrase) or a synonymous match (use of a synonym or paraphrase) in the source version is not reflected in the target version. Most important: literal or synonymous matches between stimulus and item and between a question stem and response categories. A pattern in multiple choice items is not reflected in the target version (e.g. all but one option starts with the same word, proportional length of responses options.)
Inconsistency	A recurring element across units (e.g. an instruction or prompt) is inconsistently translated, and this appears to be unintentional.
Adaptation issue	An adaptation is an intentional deviation from the source version made for cultural reasons or to conform to local usage. An adaptation issue occurs when an adaptation would be needed but was not made, or when an inappropriate or unnecessary adaptation was made.
Register / Wording issue	<i>Register</i> : difference in level of terminology (scientific term >< familiar term) or level of language (formal >< casual, standard >< idiomatic) in target versus source. <i>Wording</i> : inappropriate or less than optimal choice of vocabulary or wording in target to fluently convey the same information as in the source. This category is used typically for vague or inaccurate or not quite fluent translations.
Grammar / Syntax issue	<i>Grammar</i> : grammar mistake that could affect comprehension or equivalence, e.g. wrong subject-verb agreement, wrong case (inflected languages), wrong verb form. <i>Syntax</i> : syntax-related deviation from the source, e.g. a long (source) sentence is split into two (target) sentences or two (source) sentences are merged into a single (target) one; or another syntactic problem due e.g. to overly literal translation of the source.
Mistranslation	A wrong translation, which seriously alters the meaning. A <u>mistranslation should always be reported with a back-translation</u> . Note: a vague or inaccurate translation should rather be classified as a Register/Wording issue (or sometimes a Grammar/Syntax issue). This category covers cases where the source has been misunderstood, but also copy/paste errors that unintentionally result in a wrong text element or segment.
Guideline not followed	An explicit translation/adaptation guideline for a given text element or segment was overlooked or was not addressed in a satisfactory way.
Left in source language	A text element or segment that should have been translated was left in source language.
Minor linguistic issue	Typo or other linguistic defect (spelling, grammar, capitalization, punctuation, etc.) that does not significantly affect comprehension or equivalence. Correcting such errors is usually not controversial and can be made in track changes without documenting them.
Erratum/Update missed	An erratum or update notice has been overlooked.
Layout / Format issue	A deviation or defect in layout or formatting: disposition of text and graphics, item labels, question numbering, styles (boldface, <u>underlining</u> , <i>italics</i> , UPPERCASE), legibility of captions, tables, number formatting (decimal separators, “five” versus “5”), etc. In computer-based materials, this includes truncated words in the preview, undesired scrolling, etc.

## Annex 7.C. Translatability assessment items


Annex Table 7.C.1. Chapter 7: Translatability Assessment categories

Category	Description
Straightforward	No potential translation or adaptation problems identified during the advance translation of this segment into languages from at least two language groups.
Known difficulty, known workarounds	A translation/adaptation difficulty has been recognised in this segment and has been encountered in the past. Satisfactory solutions to this issue have been successfully implemented.
Potential issues	The current wording or content of this segment is likely to give rise to translation or adaptation problems in some languages, to the extent that functional equivalence may be difficult to achieve.
Potentially ambiguous	The current wording or content of this segment could be interpreted in more than one way and it is desirable to disambiguate the source version of this segment before submitting it for translation/adaptation.
Unnecessarily complex	The current wording or syntax of this segment is somewhat contorted, for example due to use of several clauses, questions embedded in questions or unnecessary use of passive voice. The source version can be simplified without loss of meaning.
Requires review	The current source version of this segment is not suitable for translation/adaptation and needs to be edited before submitting for translation/adaptation.
Potential cultural issue	The semantic content of this segment may be difficult to adapt in a particular cultural or language group.
Double-barrelled	A question touches upon more than one issue, yet allows only for one answer. Many double-barrelled questions can be detected by the existence of the grammatical conjunction “and” in them.
Agreement issue	There is either an agreement issue within the segment (e.g. subject-verb agreement, or sequence of tenses, or a pronoun-antecedent agreement) or an agreement issue between two segments (e.g. no grammatical match between a question and response options).
Consistency	In this segment, a different term, expression or form of address has been used versus other occurrences of similar content; and this inconsistency seems to be unintentional.
Redundancy	This segment contains a tautology or unnecessary repetition. Removing it would not alter the meaning of the segment.
Possible addition	The current wording or syntax of this segment is elliptical or unclear, and its implicit meaning is likely to get lost in translation. This could be solved by adding a word or a piece of information.
Logical problem	This segment contains a logical problem or there is a logical problem between this segment and another segment, and this issue seems to be unintentional.

## Annex 7.D. Additional items

### Annex Table 7.D.1 Chapter 7: Translation Plan 2021

Table	Title
Web Table 7.D.1	Translation Plan 2021

*StatLink*  <https://stat.link/d6xtny>

### Annex Table 7.D.2. Chapter 7: Verification outcomes regarding New Mathematics (per language)

Table	Title
Web Table 7.D.2	Verification outcomes in New Mathematics outcomes per language version

*StatLink*  <https://stat.link/9l2uf3>

### Annex Table 7.D.3. Chapter 7: Verification outcomes regarding New Mathematics (per cognitive unit)

Table	Title
Web Table 7.D.3	Verification outcomes in New Mathematics per cognitive unit

*StatLink*  <https://stat.link/k3a6rd>


### Annex Table 7.D.4. Verification outcomes regarding Creative Thinking (per language)

Table	Title
Web Table 7.D.4	Verification outcomes in Creative Thinking units per language version

*StatLink*  <https://stat.link/2pxyk5>

### Annex Table 7.D.5. Verification outcomes regarding Creative Thinking (per cognitive unit)

Table	Title
Web Table 7.D.5	Verification outcomes in Creative Thinking units per cognitive unit

*StatLink*  <https://stat.link/ai1rwx>



# 8 Field Operations

## Overview of roles and responsibilities

PISA was coordinated in each participating country/economy by a National Project Manager (NPM)<sup>1</sup> who carried out the procedures specified by the international contractors responsible for the implementation of PISA. 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 coordinated activities with school-level staff, referred to in PISA as School Coordinators.<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/economy. Major tasks carried out by the NPM included, but were not limited to:

- attending NPM meetings (in-person and virtual) and receiving training in all aspects of PISA operational procedures;
- participating in relevant webinars, such as webinars related to improving school and student participation;
- negotiating with the international contractors about local aspects of the implementation of PISA, such as national and international options, oversampling for regional comparisons, additional analyses and reporting (e.g. by language group, etc.);
- establishing procedures for maintaining the security and confidentiality of materials during all phases of the assessment implementation;
- determining the general suitability of using school computers to conduct the computer-based assessment (CBA countries/economies only) and determining the need to use laptops completely or as a supplement to school computers;
- preparing a series of sampling forms documenting sampling-related aspects of the national educational structure;
- preparing the school sampling frame and submit this to the international sampling contractor for the selection of the school sample;
- organising for the preparation of national versions of the test instruments, questionnaires, school-level materials (i.e. manuals, scripts, and forms), and coding guides;
- identifying School Coordinators from each of the sampled schools (nominated by the school principal or school staff normally responsible for testing) and working with them on school preparation activities;
- using software to select the student sample from the lists of eligible students provided by the School Coordinators;
- using software to select the teacher sample from the lists of eligible teachers provided by the School Coordinators (if applicable);

- recruiting and training Test Administrators to administer the assessments in schools;
- nominating 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 during the Main Survey only;
- monitoring the completion of School Questionnaires;
- monitoring the completion of Teacher Questionnaires (if applicable);
- monitoring the completion of Parent Questionnaires (if applicable);
- monitoring the Field Trial and Main Survey school and student participation;
- arranging for the transmission of School Questionnaire and Teacher Questionnaire (if applicable) responses completed online;
- arranging for the coding, data management, and reporting on the Parent Questionnaire (if applicable) or other national options (if applicable);
- recruiting and training coders to code the open-ended test items and the occupational data on questionnaires;
- arranging for the data entry of the test responses, Student Questionnaire responses, and School Questionnaire responses completed on hard copy in countries/economies where paper-based assessment (PBA) were administered;
- submitting the national database to the international contractor;
- submitting a written review (Field Trial Review Questionnaire and Main Survey Review Questionnaire) of PISA implementation activities after each task or 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 specific aspects of the project, such as sampling, were also provided and are described in the relevant chapters.

### ***School Coordinators***

School Coordinators were responsible for organizing school-related activities with the National Centre and the Test Administrators. A School Coordinator's Manual, prepared by the international contractors, described in detail the activities and responsibilities of the School Coordinator.

Major tasks carried out by the School Coordinator included the following:

- established the school 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;
- prepared the student list with the names of all PISA eligible students in the school and sent it to the National Centre so that the NPM could select the student sample using the ACER Maple software;
- 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 ACER Maple (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)<sup>3</sup>;

- 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);
- informed school staff, students, and parents of the nature of the assessment and the assessment date by sending a letter or organising a meeting in the school;
- secured parental permission for students to sit the assessment, if required by the school or education system;
- liaised with the Test Administrator to establish the time and other logistics of the assessment;
- informed the NPM, Test Administrator, PISA Quality Monitor 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 Coordinator was expected to ensure that the sampled students attended the assessment session(s). If necessary, the School Coordinator 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 in accordance with international standards and PISA procedures. To maintain some level of impartiality, a Test Administrator could not be the science, reading, or mathematics teacher of the students being assessed, and according to the PISA Technical Standard 8.2, it was preferred that they not be a staff member at any participating school. 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 in CBA countries/economies.

Additional responsibilities included, among others:

- ensuring receipt of the testing materials from the NPM and maintaining their security;
- contacting the School Coordinator 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 Report Form (a form designed to summarise session times, any disturbance to the session, etc.);
- in PBA countries/economies ensure that the number of test booklets and questionnaires collected from students tallied with the number sent to the school;
- in CBA countries/economies ensure that all the USB sticks used for the assessment were accounted for;
- in PBA countries/economies, collect the School Questionnaire from the School Coordinator;
- collecting Parent Questionnaires (if applicable);
- debriefing with the School Coordinator (if applicable);
- conducting a follow-up session, if needed, in consultation with the School Coordinator;

- returning the School Questionnaire, Student Questionnaires, Parent Questionnaires (if applicable), and all test materials (both used and unused) to the National Centre.

## 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 number of small schools and the homogeneity of students within schools were considered in the preparation of the school sampling plan. A school was defined as small when the approximate enrolment falls below the target cluster size. Specific details on the target population and target cluster size are presented in the sampling chapter of this technical report.

For all but one participating country/economy, 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 data file outputs for more efficient data processing later (student sampling, data analysis, etc.). 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 country/economy-specific considerations related to sampling were thoroughly documented and incorporated into the school sampling plan.

## Preparation of school-level materials

School-level materials include the School Coordinator's Manual, Test Administrator's Manual, Test Administrator's Script, the *Une Heure* (UH) Script (a national option used with Special Needs Students), and key forms (Assessment Date Form, Session Report Form, Student List, Student Tracking Form, and Worksheet for Calculating the Assessment Rate). Only English source versions of the manuals, scripts, and forms were provided by the international contractors. NPMs were required to make adaptations to these materials using the New Comment and Track Changes functions in Microsoft Word. Following approval of the adaptations, the materials were translated in the national test language(s).

In countries/economies with multiple assessment languages, the school-level materials were translated into each assessment language unless all Test Administrators and School Coordinators were multilingual. However, scripts, were required to be translated into the language of the test. After translation, the scripts underwent linguistic verification by the international contractors to ensure that they were equivalent to the source version. This verification was only done for the Field Trial. The translation of manuals and forms was not verified.

Various checking procedures were employed to review how closely national translations of the school-level materials (i.e. manuals, scripts, forms) adhered to the Technical Standards. Key elements of the adapted national language versions were reviewed in approximately 10% of countries/economies. No significant deviations were noted that might affect data validity and reliability.

## 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 Maple, the PISA student sampling software prepared by the international contractor, ACER. ACER Maple generated the Student Tracking Form (STF) which listed the sampled students for each school. The STF served as the central administration documents for the study and linked students, test booklets (PBA) or test forms (CBA), and student questionnaires. The form was also used to record student attendance (the Session Attendance Form used in prior cycles was not used for PISA 2022).

## 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/economies only);
- Student Tracking Form;
- Session Report Form;
- test delivery USB sticks (CBA countries/economies only);
- Student Login Forms (CBA countries/economies only);
- Teacher Login Forms (if applicable);
- Materials Reception Form;
- Materials Return Form;
- additional materials (e.g. COVID-19 prevention items, pens and calculators).

In PBA countries/economies, for both the Field Trial and the Main Survey, ACER Maple software pre-assigned a test booklet to each sampled student from a random starting point in each school. The software then generated the school's Student Tracking Form that 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/economies, computer-based forms were assigned automatically by the ACER Maple software based on the integrated design.

### ***Field Operations Procedures for PBA countries/economies***

The procedures 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. Two or three copies of each student's label could be printed and used to identify the test booklet and the questionnaire. Instructions were provided in the Test Administrator's Manual on how to apply labels as a quality control method to help ensure that students received the correct booklet and questionnaire. After the assessment, labels were removed and destroyed to maintain 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/economies, materials were shipped directly to the 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 such as sealing them in clear plastic or by wrapping them in paper and applying a seal. Countries/economies 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 for CBA countries/economies***

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

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

## **Test administration**

After arriving at the school on assessment day, Test Administrators were required to review the Student Tracking Form with the School Coordinator and update the form as necessary. Once the form was updated, the Test Administrator set up the room and materials for the assessment session following the steps described in the Test Administrator's Manual:

### ***Steps for setting up CBA test administration***

1. allocated a workspace and computer to each participating student.
2. set up computers for each student expected to be tested.
3. distributed Student Login Forms to students, ensuring that each student receives only the login form assigned to that student on the Student Tracking Form.
4. 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***

1. allocated a workspace to each participating student.
2. distributed test booklets (and later Student Questionnaires) to students, ensuring that each student received only the test booklet assigned on the Student Tracking Form.
3. wrote the testing date on a board or sheet of paper visible to all students.
4. asked the students to write the test date on their test booklet covers (and later the Student Questionnaire).
5. 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 strictly follow the timing of the paper-based assessment, especially the administration of the test sessions (2 sessions of exactly 1 hour each). The timings were the same for CBA test sessions, with additional time added if one or more of the optional questionnaires was administered. Although CBA test sessions were timed by the student delivery system, Test Administrators were still required to enforce the timing and not move students forward prematurely. The timing of the is shown in Table 8.1. below.

**Table 8.1. Timing of the CBA and PBA assessment sessions**

Activity	Timing
Distributing materials and reviewing general directions	15 minutes (approximately)
First 60 minutes of test	60 minutes (exactly)
Short break	<b>Generally, no more than 5 minutes</b>
Second 60 minutes of test	60 minutes (exactly)
Break	<b>15 minutes*</b>
Student Questionnaire	35 minutes (approximately) + additional time for any optional questionnaires
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 strict. The recommended amount of time is 15 to 30 minutes, but the time can be adjusted at the discretion of the National Centre, and school's circumstances.

NPMs were allowed to adapt the length of the short break between the two testing sessions. Most countries/economies allowed only the recommended 5-minute break. In a few cases, countries/economies did not offer a break between test sections in all of their schools as they felt this would be too disruptive. Some countries/economies required a longer break usually up to 15 minutes.

No changes to the timing of the test sessions were allowed. Adaptation to the timing of the Student Questionnaire session (for both CBA and PBA) was possible in order to allow students to finish answering the questionnaires and maximise the contextual data obtained from students. If a few students were still working at the end of the allotted time for the questionnaire session, 10 additional minutes were given to allow completing it.

The test scripts for both CBA and PBA sessions had to be read to the students word-for-word to maintain standardised assessment procedures across all participating countries/economies. For PBA sessions, the Test Administrators were required to read the practice exercises and other key instructions to the students. Therefore, if a student 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 assessment introduction, the Test Administrators informed the student about the purpose of the test and would allow the student to begin.

For both CBA and PBA sessions, students were not allowed to leave during 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 session (CBA sessions) or collect the student's test material (PBA sessions). If the student was present for any part of the assessment, they were recorded as participating even if they did no work at all.

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

Observers during the testing sessions were generally limited to necessary staff members and the international PISA Quality Monitors. National Centre staff were encouraged to observe assessments when possible. National Centres were responsible for ensuring that confidentiality arrangements were in place. In most cases, it was national policy to require observers to sign a confidentiality agreement.

At the end of the computer-based administration (cognitive 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 (or returned to the National Centre) all login forms. The Test Administrator then collected all USB sticks (if used) and conducted a quality-control check on the number of USB sticks and the information on the Student Tracking Form 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 Student Tracking Form, and Session Report Form and shipped to the National Centre, typically within 24 hours of completing the assessment, or the follow-up session.

At the end of the paper-based administration, Test Administrators collected all assessment materials and the completed School Questionnaire from the School Coordinator. The assessment material from each administration session were bundled together with the corresponding Student Tracking Form, Session Report Form, unused test booklets, and Student Questionnaires. These were shipped to the National Centre, typically within 24 hours of completing the assessment or follow-up session.

Any missing secure and confidential material had to be reported to the Survey Operations team at Westat and to the National Centre as soon as possible, and no later than 24 hours after the discovery of the missing data. National Centres are asked to use a standard form to report missing items and what was done to recover them.

### Receipt of materials at the national centre after testing

The procedures recommended that the National Centre establish a database of sampled 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/economies).

The procedures also 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.

### Field Trial and Main Survey reviews

NPMs were required to complete a structured review of their Field Trial and Main Survey operations. These were submitted via SurveyMonkey (an online survey platform) preferably on an on-going basis after the completion of each activity. The complete review questionnaire was due 4 weeks after the submission of the national database.

These reviews were an opportunity to provide feedback to the National Centres, international contractors, and the OECD on the various aspects of the implementation of PISA and to provide suggestions for areas that could be improved either for the Main Survey or for future cycles.

The data from these two questionnaires were compiled into reports, which were released after the Field Trial and after the Main Survey.



## Notes

---

1. Some participating countries/economies had more than one National Project Manager.
2. Throughout this document, the terms “School Coordinator” and “Test Administrator” are used when discussing the administration of the test in schools. However, please note that some countries/economies use the term School Associates. These are individuals who simultaneously fulfil the role of both School Coordinator and Test Administrator. School Associates received a School Associate’s Manual and were trained by the National Centre. For the sake of simplicity, we do not refer to School Associates specifically in the text.
3. Some participating countries/economies chose to use the Une Heure (UH) option, which is a 1-hour version of the PISA assessment meant for students who are considered unable to take the full PISA assessment. These students were assessed in separate sessions. Some countries/economies also provide other PISA-approved accommodations.

# 9 PISA Quality Monitoring

## Introduction

PISA data collection activities were undertaken in accordance with strict quality assurance procedures. These procedures have two components: first, to develop and document procedures for data collection; and second, to monitor and record the implementation of those procedures. Chapter 8 describes the procedures which National Centres were required to follow while this chapter considers the second part of the process – monitoring data collection quality.

While the aim of quality control was 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 Consultations,
- PISA quality monitor (PQM) Hiring Process,
- PISA quality monitor training,
- PISA quality monitor visits for the Main Survey,
- Data adjudication.

## 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 cycles of the PISA assessment at both the jurisdiction and international levels.

The Field Trial Review and the Main Survey Review Questionnaires were submitted via SurveyMonkey (a secure online survey platform). The review questionnaires were due no later than 4 weeks after the submission of the national database, which in turn is due no later than 8 weeks after the last date of testing, or on a flow basis after completion of each phase such as translation of instruments. The data from these two questionnaires were compiled into reports that were released after the Field Trial and Main Survey.

The Field Trial and Main Survey Review Questionnaires were organised around the different activities that took place during the Field Trial and Main Survey phases of the assessment. A rating system was used to document NPMs' level of satisfaction with or comments on:

- use and clarity of key documents and processes;
- communication with the international contractors;
- review of the quality of communication by activity;
- review of the usefulness of the PISA Portal;

- review of the quality and usefulness of the meetings (in person and virtual);
- breaches of security and/or confidentiality;
- review the sampling tasks, the sampling software (ACER Maple) and the sampling process;
- review the translation, adaptation and verification processes;
- preparation of school-level materials and the process for adapting them, the webinars given on Test Administrator (TA) training and gaining co-operation and other test administration procedures;
- review of the coding process including coder training, coding systems and coding occupational categories; and
- review the data management process including data entry, data importing, data submission and data cleaning.

## National centre consultations

Constant consultations took place between senior international contractor staff, NPMs or other representatives of National Centres throughout the entire PISA 2022 cycle. The consultations provided the opportunity for detailed discussions on a wide variety of PISA implementation questions and concerns.

## PISA Quality Monitor Hiring Process

The number of PQM hired depended on the specific situation in each jurisdiction. For jurisdictions with a six to eight week assessment period, three PQMs generally were required. Shorter assessments required more PQMs. Jurisdictions with adjudicated regions usually required more PQMs. The number of PQMs per jurisdiction for PISA 2022 ranged from one to eight.

All PISA Quality Monitors were nominated by the NPMs and sent to the international survey operations contractor. Based upon the NPM nominations, which were usually accompanied by candidate CVs, the survey operations contractor selected monitors who were independent from the National Centre, generally knowledgeable in testing procedures or with a background in education and research and able to communicate adequately in English. In this context, independent from the National Centre means: a) not paid by or reporting directly to the NPM, b) not an immediate familiar member of the NPM or National Centre staff.

Suitable candidates were further vetted by the international survey operations contractor who interviewed them usually remotely. In the case of candidates returning from the PISA 2018 cycle, they received updated information via emails and sometimes were contacted by Zoom or WhatsApp if there were further questions. The survey operations contractor was responsible for hiring candidates in each of the participating jurisdictions, organising their training, selecting the schools to visit and collecting information from the PQM visits. Before getting access to confidential material such as the names of participating schools, names of students or test material, every PQM signs an Honoraria and Confidentiality Agreement.

## PISA Quality Monitor Training

After signing the Honoraria and Confidentiality Agreement, PQMs also were given access to the school-level materials (manuals and script in both English and the regional language).

Each PQM was required to participate in two trainings: The National Centre Test Administrator Training and the PQM online training presented by the survey operations contractor The Test Administrator Training

was in-person, online, a combination of the two, or self-study. The purpose of this training was to familiarise the PQM with the tasks and procedures TAs needed to successfully conduct assessments.

Prior to the PQM training, PQMs received the PQM Manual and the Data Collection Form (DCF) used to document assessment observations. This training reviewed their role and responsibilities as quality monitors and familiarised PQMs with general PISA procedures and policies. After training, PQMs were required to complete a quiz that was reviewed by survey operations staff who provided feedback as needed. Survey operations contractor staff continued to be available to the PQMs when updates were needed or they had any questions or concerns.

## PISA Quality Monitor Visits

PQMs visited a subset of schools to observe and to document the test administration. In each jurisdiction, at least, 15 schools (or sessions if more than one session was observed in a school). Five schools at a minimum were observed in each adjudicated region.

Survey operations contractor staff worked with each PQM to develop a schedule of school site visits to ensure that a range of different schools (roughly corresponding to the sampling strata plan) was covered and that the schedule of visits was both economically and practically feasible. Upon completion of their observations, the international survey operations contractor paid approved expenses and fees directly to each monitor.

Prior to visiting a school, PQMs contacted the School Coordinator and/or school principal to explain the purpose of the visit and to obtain information about the arrival time and other logistical information about the visits. Test Administrators were not informed of these visits in advance. School Associates who served as both TA and School Coordinators (SC) were informed of PQM visits in advance.

The international survey operations contractor also provided support to the National Centres throughout the data collection phase and addressed any issues or concerns with National Centres that were noted during the quality monitor visits.

### ***Information collected in PQM observations***

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

- preparation for the test session,
- testing environment,
- conducting the assessment
  - session date and timing
  - deviations from standard test procedures
  - conduct of the students,
- administering the questionnaire,
- other comments about the test session.

PQMs recorded all key test session information using a hard copy of the DCF. After each session, the monitor entered the data into the SurveyMonkey form.

This information was used to check that the implementation in each session was in accordance with the PISA Technical Standards. Discrepancies were reported to National Centres and clarified as needed. The information was also called upon if other contractors or the Technical Advisory Group (TAG) had any concerns or questions about the data and data collection process as mentioned below.

## 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 TAG to consider during the data adjudication process.

The TAG experts used the quality-monitoring reports from the central data adjudication database to make individual evaluations for each jurisdiction on the quality of school and student sampling, survey operations, translation and coding and data quality. The final reports by TAG experts were then used for the purpose of data adjudication that took place prior to the release of the data in 2023.

# 10 Survey Weighting and the Calculation of Sampling Variance

Survey weights are required to analyse PISA data, to calculate appropriate estimates of population parameters, their 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 population parameters and of standard errors, and to conduct significance tests and create confidence intervals appropriately, taking into account the complex sample design used to select individual student participants for PISA.

## Survey weighting

While the students included in the final PISA sample for a given country/economy were chosen randomly, the selection probabilities of the students vary. Survey weights must be incorporated into the analysis to ensure that each participating student appropriately represents the correct number of students in the full PISA population. Sampling weights are used to control the proportional contribution of each participating unit to the overall population estimate.

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

- 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. 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 using survey weights.
- 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 smaller than expected, a larger proportion of 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/economy. 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 may be effective in reducing standard errors (Kish, 1992<sup>[1]</sup>).
- In countries/economies that opted to participate in the financial literacy study, additional students were selected in all schools. Since the financial literacy sample was also designed to represent the full PISA student population, the weights for the sampled students were adjusted to account for this. Different adjustment factors applied to each student's weight, depending on which assessment form the student was assigned.

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 are used in other international studies of educational achievement such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress of International Literacy study (PIRLS), among others. The underlying statistical theory for the analysis of survey data can be found in Cochran (1977<sup>[2]</sup>), Lohr (2010<sup>[3]</sup>) and Särndal, Swensson and Wretman (1992<sup>[4]</sup>).

Weights are generally 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 four adjustment factors, and can be expressed as:

$$W_{ij} = \{[(w_{1i} * t_{1i}) * f_{1i}] * (w_{2ij} * f_{2ij})\} * t_{2ij}$$

Formula 10.1

Where:

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

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

$f_{1i}$  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);

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

$f_{2ij}$  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; 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 calculated as:

$$w_{1i} = \begin{cases} I_g / MOS_i & \text{if } MOS_i < I_g \\ 1 & \text{otherwise} \end{cases} \quad \text{Formula 10.2}$$

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/economies that did a computer-based assessment, and 35 for most countries/economies that did 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:

$$\begin{aligned} MOS_i &= EST_i && \text{if } EST_i \geq TCS; \\ &= TCS && \text{if } TCS > EST_i \geq TCS/2; \\ &= TCS/2 && \text{if } TCS/2 > EST_i > 2; \\ &= TCS/4 && \text{if } EST_i = 0, 1 \text{ or } 2. \end{aligned} \quad \text{Formula 10.3}$$

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 100 15-year-old students at the time of sample selection then  $MOS_i = 100$ . And, if the country/economy 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.

In the case of replacements, the  $MOS_i$  used in the calculation of the school base weight is that of the replacement school (not the original school).



### ***The school base weight trimming factor***

Once school base weights were established for each sampled school in the participating country/economy, 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.

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 where the 15-year-old student enrolment exceeded  $3 \times \text{MAX}(TCS, MOS_i)$  were flagged. 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  $3 \times \text{MAX}(TCS, ENR_i)$  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, with the aim of not introducing any substantial bias by altering many student weights to a large degree. The school trimming happened in 13 participating countries/economies. There were four school weights trimmed for Cambodia and Panama respectively, and six school weights trimmed for Denmark. In the remaining countries/economies where some trimming was needed only one or two school weights were trimmed.

### ***The school non-response adjustment***

In order to adjust for the fact that those schools that declined to participate, and were not replaced, were not in general typical of the schools in the sample as a whole, school-level non-response adjustments were made. Within each participating country/economy 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 non-participating schools and their students.

The compositions of the non-response groups varied among countries/economies, but the original adjustment groups for all countries/economies were formed by cross-classifying the explicit and implicit stratification variables used for school sample selection. Usually, about 10 to 40 such groups were formed within a given country/economy depending upon school distribution with respect to stratification variables. If a country/economy 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 groups where all schools participated, as the school non-response adjustment factor was 1.0 regardless of whether groups were collapsed or not. However, since the groups used for school non-response adjustment were also used as the basis for student non-response adjustment, such groups 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, groups were generally collapsed starting from the last implicit stratification variable until the violations no longer existed. In countries/economies with very high overall levels of school non-response after school replacement, explicit strata were sometimes collapsed.

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

$$f_{1i} = \frac{\sum_{k \in \%OMEGA(i)} w_{1k} enr(k)}{\sum_{k \in \Gamma(i)} w_{1k} enr(k)} \quad \text{Formula 10.4}$$

where  $enr(k)$  is the actual enrolment of 15-year-old students in the school at the time of preparation of the student list (and so, in general, is somewhat different from the  $EST_i$ ), 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.

Annex Table 10.A.2 shows the number of school non-response classes that were formed for each country/economy, and the variables that were used to create the cells.

### **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:

$$w_{2ij} = enr_i / sam_i \quad \text{Formula 10.5}$$

where  $sam_i$  is the number of students sampled 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 a proportion of students in the school.

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/economies with extra direct-sampled grade students (e.g., 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 France and Germany. The extra weight component consisted of the class weight for 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 within-school non-response adjustment**

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

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

Formula 10.6

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 participating country/economy were defined so that each grade category contained a substantial proportion of the PISA population in each original explicit stratum or final school non-response adjustment groups where collapsing crossed explicit strata. The definition was then applied to all schools in the same original explicit stratum or in the same final school non-response adjustment group.

In most cases, the 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 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, adjustments greater than 2.0 were flagged for review, 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 participating countries/economies 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 33%, the school was treated as a non-respondent, and its student data were removed.

For countries/economies with extra PISA immigrant student (Denmark, Finland) or 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 students. No procedural changes were needed for France and Germany 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 original explicit stratum. The sample design was intended to give all students from within the same original 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 relatively large weights, which increases the sampling variance. The student non-response adjusted 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 happened in about 11% of all participating countries/economies.

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 within each explicit stratum, 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***

Spain had a financial literacy subsample of its national sample, which required a separate weighting stream. The extra weighting stream followed all the usual weighting steps.

A few other countries/economies also had national option students but, in these cases, weighting was done along with the PISA students (i.e., Denmark, Finland, and Iceland) if weights were required. Specifics about national options are beyond the scope of this report.

### ***International options***

For the teacher questionnaire (TQ), special weight factors were applied at the end of weighting in 18 countries/economies to ensure that in the TQ database, the sum of weights of the math and non-math teachers would still approximate the math and non-math teacher population, respectively. For financial literacy, special weight factors were applied at the end of weighting to ensure that in the financial literacy database, the sum of weights of the financial literacy students would still approximate the PISA population. The overall, math, and non-math weighted teacher questionnaire response rates were calculated. The weighted financial literacy response rates were also calculated.

## **Teacher weighting**

While the TQ has been an international option in past cycles, the PISA 2022 cycle is the first cycle in which survey weights were calculated for sampled teachers. This section describes the methodology for calculating teacher weights. Eighteen countries/economies participated in the TQ option. Teachers eligible for TQ were those that were currently teaching the modal grade(s) of PISA-eligible students in the country/economy. In 2022, the TQ option consisted of separate samples of mathematics teachers and 'other' teachers (those not teaching mathematics).

It is possible that a teacher who was identified as a mathematics teacher on the teacher list provided by the school was found to be a non-mathematics teacher based on their response in the TQ, and vice versa. On the rare occasions that this occurred, the teacher weight was calculated based on their classification at the time of selection (i.e., as identified on the teacher list). In the delivery file, the teacher 'type' (mathematics or non-mathematics teacher) identified on the teacher list and the teacher 'type' identified by the teacher in their TQ are both available for analysis purposes.

The TQ weighting methodology followed closely the approach described in the previous section for student weighting. However, there are several differences, and these are described in the subsections that follow.

### ***The TQ school base weight***

Because TQ data were collected primarily for use in conjunction with the data of participating PISA students, the set of participating schools identified during student weighting was determined to be the set of participating schools for teacher weighting. Therefore, any responding teachers outside of these schools were dropped from the TQ sample. The final school weights from student sampling were used for calculating TQ weights. These final school weights incorporate school base weight trimming and school non-response adjustments, and these are described in some detail earlier in this chapter.

It is possible that a participating school did not have a teacher list completed and, as a result, had no teachers sampled. Such schools will usually be ineligible for the TQ, because they would have no PISA-eligible students in the modal grade. However, it is possible that a TQ-eligible school had no teachers listed or sampled. For such schools, an additional school nonresponse stage was carried out. There were five countries for which this extra adjustment was required, with the number of schools shown in parentheses – Australia (3), Brazil (5), Colombia (2), Hong Kong (5), and Panama (5). These schools were coded as nonrespondents for the purpose of TQ weighting, and the final school weights of other participating schools in the same final school nonresponse adjustment cell from student weighting were increased to account for this additional school nonresponse.

Where the teacher response rate within a participating school was low (or 0%), this was handled through teacher nonresponse adjustment. A school-level teacher participation rate was calculated and included as a variable on the teacher delivery file. This information can be used as a measure to provide data users with information about the quality of school-level TQ data.

### ***The within-school teacher base weight***

The within-school teacher base weight was calculated in the same way as the within-school student base weight. Since the samples of mathematics teachers and non-mathematics teachers are selected independently, teacher weights for mathematics teachers within a particular school will differ from weights for non-mathematics teachers. However, within a particular school, all mathematics teachers have the same within-school base weight, and all non-mathematics teachers have the same within-school base weight. The formula for within-school teacher base weights can be written as follows:

$$w_{2ikl} = enr_{ik}/sam_{ik}$$

Formula 10.7

where  $k=1$  or  $2$ , to indicate mathematics or non-mathematics teachers,  $enr_{i1}$  and  $sam_{i1}$  are the number of mathematics teachers and *sampled* mathematics teachers respectively in school  $i$ , and  $enr_{i2}$  and  $sam_{i2}$  are the number of non-mathematics teachers and *sampled* non-mathematics teachers respectively in school  $i$ .

### ***The within-school teacher non-response adjustment***

The teacher nonresponse adjustment followed the same approach as the student nonresponse adjustment. For teachers, the only information available besides the final school nonresponse cell and explicit stratum is the teacher type (mathematics or non-mathematics teacher). Within each final school non-response adjustment cell, explicit stratum, and teacher type, and school combination, the teacher non-response adjustment  $f_{2i}$  was calculated as:

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

Formula 10.8

where,

$\Delta(i)$  is all participating teachers in the final school non-response adjustment cell and explicit stratum-teacher type-school combination; and,

$X(i)$  is all participating and non-participating teachers in the final school non-response adjustment cell and explicit stratum-teacher type-school combination. Ineligible teachers are excluded from the calculation. Note that there no *excluded* teachers.

Collapsing of teacher non-response adjustment cells was done as needed to ensure at least 15 participating teachers were in each final adjustment cell. Because the number of sampled mathematics teachers in each school was often small, collapsing across schools was always required for mathematics teachers, and there were instances where it was necessary to collapse teacher types.

### ***Trimming the teacher weights***

The PISA sample design is intended to produce a self-weighting sample of *students*, in the absence of school and student nonresponse. There are several reasons why final student weights vary, and these are described at the beginning of this chapter. However, extreme outlier student weights are typically due to poor frame data on school-level student enrolment. As described in the student weight trimming section, extreme student weights are trimmed in order to reduce the sampling variance.

In contrast, the PISA sample design was not intended to produce self-weighting samples of teachers. Schools were sampled proportional to student enrolment, and while the number of mathematics and non-mathematics teachers in a school can be expected to be correlated with student enrolment, this relationship varies from school to school, and no steps were taken to reduce the weight variability of the teacher samples. Since teacher weights vary considerably *by design*, there was no clear basis to identify ‘outlier’ teacher weights. It was decided that no trimming of teacher weights would be carried out.

## **Calculating sampling variance**

A replication methodology is employed to estimate the sampling variances of the PISA parameter estimates. This methodology accounts for the variance in estimates due to the sampling of schools and students. Additional variance due to the use of plausible values drawn from the posterior distributions of scaled scores is captured separately as measurement or imputation error. Computationally the calculation of these two components could be carried out using a single program, such as *WesVar 5*, or with the IDB Analyzer using R, SPSS and SAS macros developed for this purpose.

### ***The balanced repeated replication variance estimator***

The specific replication 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 PIRLS, and it is well documented in the survey sampling literature [see Rust (1985<sup>[5]</sup>); Rust and Rao (1996<sup>[6]</sup>); Rao and Shao (1996<sup>[7]</sup>); Wolter (2007<sup>[8]</sup>)]. 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, and for which the jackknife methods 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<sup>[9]</sup>).

For a country/economy where the student sample was selected from a sample of schools, rather than all schools, the BRR method was implemented 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. In the case of an odd number of schools within a stratum, a triplet 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, variance 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 triplet), 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 school 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 the same order. Details concerning Hadamard matrices are given in Wolter (2007<sup>[8]</sup>). 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<sup>[10]</sup>).
- To use a Hadamard matrix of order 80 requires that there be no more than 80 variance strata within a country/economy, 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/economies, 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/economies for part of the sample (and for the entire samples for Brunei, Iceland, Macao (China), Malta, North Macedonia, and Qatar), 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, there could have been a stage of sampling that precedes the selection of schools. Then the procedure for assigning variance strata, variance units and replicate factors would be applied at this higher level of sampling. The schools and students would then inherit the assignment from the higher-level unit in which they were located. No countries/economies used such a three-stage design for PISA 2022.
- Procedural changes were in general not needed in the formation of variance strata for countries/economies 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 France and Germany, since a separate weighting stream was needed in these cases.

The variance estimator for the BRR method is then calculated using the following formula:

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

Formula 10.9

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.

### ***Reflecting weighting adjustments***

Implementing this approach required that the PISA Consortium produce a set of replicate weights in addition to the full sample weight. 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 trimming the school base weight, re-computing the school non-response adjustment for each replicate, applying the adjustment for student selection (the student base weight component), computing the student non-response adjustment for the replicate, and trimming the student non-response adjusted weight. The school and student non-response adjustments were recalculated and applied to each set of replicate weights using the methodology described earlier in this chapter. Like the full-sample adjusted student weight, the replicate adjusted student weights are provided as variables in the data file.

To estimate sampling errors correctly, the analyst must use the variance estimation formula above, by deriving estimates using the  $t^{\text{th}}$  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/economy, but the procedure gives a more accurate estimate of sampling variance.



### **Countries/economies where all students were selected for PISA**

In Brunei, Iceland, Macao (China), and Malta, 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/economies, 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/economy there was some student non-response. Not all PISA-eligible students were assessed, resulting in 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 of the 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.

### **Variance estimation for the TQ sample**

The TQ sample used the same variance estimation approach as the student sample. Since the participating schools for the student sample were used as the participating schools for the teacher sample, the full sample final school weight for the student sample was also the full sample final school weight for the teacher sample. Similarly, the replicate school weights for the student sample were used as the replicate school weights for the teacher sample. For certainty schools, instead of pairing schools, pairs of individual teachers were formed from within the same school and the procedure of assigning variance units and replicate weight factors was then conducted at the teacher level, rather than at the school level. Teachers were sorted by teacher type before pairing was done, to maximise the chance of pairing teachers of the same teacher type.

## **References**

- Adams, R. and M. Wu (eds.) (2002), *Programme for International Student Assessment (PISA): PISA 2000 Technical Report*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264199521-en>. [10]
- Cochran, W. (1977), *Sampling Techniques (3rd ed.)*, John Wiley and Sons. [2]
- Judkins, D. (1990), "Fay's method for variance estimation", *Journal of Official Statistics*, Vol. 6, pp. 223-229. [9]
- Kish, L. (1992), "Weighting for unequal Pi", *Journal of Official Statistics*, Vol. 8/2, pp. 183-200. [1]
- Lohr, S. (2010), *Sampling: Design and Analysis*, Brooks/Cole, Boston, MA. [3]
- Rao, J. 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. [7]
- Rust, K. (1985), "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics*, Vol. 1/4, pp. 381-397. [5]
- Rust, K. and J. 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. [6]
- Särndal, C., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York, NY. [4]

Wolter, K. (2007), *Introduction to Variance Estimation*, Springer, New York, NY.

[8]

# Annex 10.A. School non-response items

## Annex Table 10.A.1. Chapter 10: School non-responses

Tables	Title
Table 10.A.2	School non-response classes

## Annex Table 10.A.2. School non-response classes

Country/Economy	Number of explicit strata*	Implicit stratification variables	Number of original cells	Number of final cells
Albania	12	ISCED level (3), Gender (5)	70	16
Argentina	21	Department (19); Location (2); Level (8); Performance (5)	193	43
Australia	25	Geographic Location (3); School gender composition (3); School socio-economic Level (11); ISCED level (3)	385	77
Austria	18	Region (9); Percentage of girls (5); Programme for Statut schools (3)	276	29
Baku (Azerbaijan)	5	None	15	11
Belgium	31	School type – French Community (4), German and Flemish Community (1); 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)	164	32
Brazil	20	State (27); ISCED level (5); Urbanisation (2); Capital/Country (2); IDH Quintiles (5); School gender composition (3)	506	63
Brunei	8	Sixth Form (3); District (4)	17	5
Bulgaria	3	Type of school (3)	9	9
Cambodia	18	School management (2); Shifts (2)	40	23
Canada	67	Urbanicity (2); Funding (2); ISCED Level (3)	208	38
Chile	14	School Type (4); National test score level (4); Percentage of girls (6); Urbanicity (2); Geographic	177	28

		zone (4)		
Chinese Taipei	19	Funding (2); Region (6); School gender composition (3); Municipality (2); Shift offerings (2)	141	35
Colombia	6	Regional entities (96); Main shift (2); School gender composition (5)	176	33
Costa Rica	6	Zone (2); Track (2); Shift (2); Education regions (27); ISCED level (3)	112	34
Croatia	7	Region (6); School gender composition (3)	56	20
Cyprus	8	Urbanisation (2); Language (2)	16	9
Czech Republic	32	Region for school types 3, 4, 5 (14); Gender (3)	146	37
Denmark	6	School type (7); ISCED level (3); Urbanisation (5); Region (5); FO group (3)	152	42
Dominican Republic	10	Shift (6); School size (4); Programme (4)	88	23
El Salvador	28	Founding (2); ISCED level (3); Study Commitment (3)	107	26
Estonia	4	School type (3); Urbanicity (2); County (15); Funding (2)	71	15
Finland	30	Immigrant cluster (6); Regional State Administrative agencies (7); School type (5)	62	24
France	22	Secteur (2)	32	14
Georgia	9	Language (9)	22	9
Germany	18	State for SEN and vocational schools only (16); School type for Normal schools (6)	68	24
Greece	3	Funding and region (15); School type (4)	100	26
Guatemala	8	ISCED (2); Modality of teaching (4)	25	11
Hong Kong (China)	5	Student academic intake (4); School gender composition (3)	21	8
Hungary	6	Geographical region of Hungary (7); Average mathematics performance in the National ABC 2020 (6)	132	49
Iceland	24	Urbanicity (2)	23	10
Indonesia	4	School type (5); Funding (2); Region (8)	95	48
Ireland	9	School Gender Composition (4); Socio-Economic Status Quartile (4)	73	21
Israel	13	ISCED level (3); Group size (2); Socio-Economic Status (3);	71	26

		Geographic/Administrative District (2)		
Italy	36	Region (20); Types of School (2)	107	32
Jamaica	15	Gender (3); School types (5)	41	15
Japan	4	Levels of proportion of students taking University/College Entrance Exams (4)	14	9
Jordan	8	Region (3); Urbanisation (2); School gender composition (3); Level (2); Shift (2)	96	36
Kazakhstan	19	ISCED level (2); Location (2); Language (3); Funding (2); Shifts (2)	190	69
Korea	6	Urbanisation (3); School gender composition (3)	26	15
Kosovo	8	Urbanisation (2); ISCED (3)	26	11
Latvia	4	School type (4)	15	11
Lithuania	21	School language 2 (4); School location (5); School type (5); School type 2 (2)	45	18
Macao (China)	10	Gender (3); School orientation (2)	18	9
Malaysia	10	School type (18); Location (2); Gender (3); ISCED level (2)	32	11
Malta	3	N/A	9	7
Mexico	12	School program (8); Urbanisation (2)	45	15
Mongolia	16	Property type (3); ISCED orientation (2); ISCED level (3)	27	14
Montenegro	12	Gender (3)	19	15
Morocco	12	Milieu (2); Type (2)	31	22
Netherlands	10	N/A	28	10
New Zealand	4	School decile (4); School authority (2); School gender composition (3); Urbanicity (2)	41	14
North Macedonia	9	Urbanisation (2)	14	9
Norway	2	None	6	3
Palestinian Authority	7	Region (2); Gender (3); District (25)	121	35
Panama	16	Educational region (16); ISCED level (3); Programme orientation (4); Language of test (3)	98	18
Paraguay	19	Region (5)	66	20
Peru	4	Region (26); School gender composition (3); School type (4)	107	30
Philippines	16	School Management (2); Type of Community (3); ISCED Level (3); Gender Composition (5)	73	24

Poland	4	Private/Public (2); Locality size (4); School gender composition (3)	41	6
Portugal	26	ISCED (3); School management (2); School Location (3); Curriculum (3)	97	35
Qatar	4	Level (5); School gender composition (3); Language (2); Programme orientation (3)	39	13
Republic of Moldova	28	Funding (2); ISCED program orientation (6)	38	14
Romania	6	School location area (2); Development regions (8)	46	18
Saudi Arabia	30	Education District (47); School Level (2)	104	37
Serbia	22	Region implicit (5); School type implicit (7); Language (2)	45	25
Singapore	4	Gender (3)	5	4
Slovak Republic	24	T9 - Three-year average of scores in national testing in math and Slovak (Hungarian) language (7); School type (6); Language (3); Funding (3)	146	32
Slovenia	7	Location (5); School Gender Composition (3)	149	33
Spain	40	Linguistic model – for Basque Country only (3), other regions (1)	121	100
Sweden	8	Geographic LAN for upper secondary only (21); Responsible authority, if upper secondary (3); Percentage of immigrant students (3); Income quartiles, if ISCED2 (4)	65	21
Switzerland	15	Sponsorship (2); School type (33); Canton (30); Foreign Speaking Student Share (3)	197	32
Thailand	15	Public/Private (2); Region (9); Urbanisation (2); School gender composition (3)	135	33
Türkiye	36	Statistical Region Unit (12); Location (2); Gender (3)	191	27
Ukraine (18 of 27 Regions)	49	ISCED Orientation (3); Language (3)	87	23
United Arab Emirates	47	School gender (3); Language (3); ISCED (3); Programme (2)	146	73
United Kingdom (excl. Scotland)	34	Gender (3); School performance – England (6) and Wales (5); Local authority (7)	332	47
United Kingdom (Scotland)	8	Gender (3); Area type (6)	32	13

United States of America	8	Grade span (5); Urbanisation (4); Minority status (2); Gender (3); State (51)	210	20
Uruguay	11	Location/Urbanisation (4); School gender composition (4)	40	16
Uzbekistan	27	Specialization (2)	49	19
Viet Nam	15	Region (6); Province (63); School type (4); Study commitment (2)	157	29

# 11

## Scaling PISA data

### Overview

The test design for PISA 2022 follows the balanced incomplete block (BIB) design used in prior cycles, with adaptations to incorporate multi-stage adaptive testing (MSAT) for the reading and mathematics domains. With the traditional BIB design, units (i.e., small sets of items) are grouped into mutually exclusive clusters (i.e., sets of units) assembled into test forms. For the non-adaptive domains, the clusters are distributed so that they appear with equal frequency across forms and positions within forms, which leads to the design being balanced. When these tests are administered, students are administered a randomly selected test form so that differences in the average test performance on forms consisting of different sets of items are not due to differences in student proficiency. However, the test forms can be of different difficulty, which means that the performance of groups measured through different sets of items cannot be directly compared using total-score statistics such as the average number or percent of items that the student responded to correctly.

The limitations of using the number or percent of items correct to score assessments that are designed with BIB or administered through MSAT can be overcome by modelling the item responses through item response theory (IRT). When students respond to a set of items in a common subject or domain, their response patterns should show regularities that can be modelled using the underlying commonalities among the items. This regularity can be used to characterize the students and items on a common scale, even when students take different sets of items. However, IRT is only the first step in the scaling of PISA data that makes it possible to describe the distributions of student performance in populations or subpopulations, to estimate the relationships between proficiency and background variables, and to build and select test forms that match the difficulty of the form with the ability of students.

The scaling approach employed in the analyses of PISA data (*population modelling*) combines IRT and latent regression modelling to increase overall measurement accuracy and to avoid potential bias in the estimation of the relationships between proficiency and contextual variables from the background questionnaire (BQ). Once the population model is estimated, multiple plausible values can be drawn for each student from a posterior distribution of proficiency that accounts for the sources of uncertainty in the data.

In PISA 2022, mathematics and reading MSAT designs were incorporated into the overall BIB design to deliver a 60-minute MSAT to students, instead of the two 30-minute clusters used for the other domains. The reading design was the same that was used in 2018. However, as reading became a minor domain, some of the items were released and the 2018 testlets that lost some items were re-assembled from the reduced item pool in a way that minimized the changes. As in 2018, the reading design included a proportion of student misrouted from the core to stage and from stage 1 to stage 2 to ensure that responses on all items were collected from students across a broad proficiency range. The reading design partially balanced item position between stage 1 and stage 2. For mathematics, a newer design was implemented that fully balanced item position across core, stage 1 and stage 2 and randomly assigned 25% of the students to a linear design to ensure that item responses are collected from students across a broad proficiency range (for further details, see Chapter 2 in this report).



However, despite these design differences across domains, for the most part, the same classical analysis (item analysis - IA and timing), item response theory (IRT) and population modelling procedures could and were effectively implemented to fulfil all the main survey analyses goals.

This chapter first describes the quantity and quality of the data submitted by the participating countries/economies. Analyses were conducted to evaluate how well the assessment design was reflected in the data and to verify that the data quality was appropriate for IRT and population modelling. The subsequent sections explain the models and methods used for IRT, latent regression modelling, and the generation of plausible values. Then, the application of these models and methods to the PISA 2022 data to produce the national and international item parameters and the plausible values are described. Finally, the approach and methods used for estimating the linking errors between the 2022 main survey and the previous PISA cycles are explained.

## Data yield and data quality

Before the data were used for scaling and population modelling, analyses were carried out to examine the quality of the data to ensure that the test design requirements were met, and also to verify that the data reflected the intended design. The following subsections give an overview of these analyses and their results. Overall, the quality of the data and the cognitive instruments met the requirements for the intended analyses and scaling methods. The results of the item analyses were communicated to countries/economies for their review and feedback. Taken together, the data yield and item analyses confirmed that the PISA 2022 computer platform had successfully delivered, captured, and exported the student- and item-level data expected from both the computer-based assessment (CBA) and paper-based assessment (PBA).

### **Target sample size, routing, and data yield**

#### *Target sample size*

The assessment design for the PISA 2022 main survey included the core domains of reading, mathematics, and science, delivered through both CBA and PBA. In addition, it also included the optional domain of financial literacy and the innovative domain of creative thinking, both delivered only through CBA. As part of the sampling design, participating countries/economies were required to sample a minimum of 150 schools to cover their national population of 15-year-old students. Countries/economies taking the CBA with creative thinking (CrT) or the CBA without CrT needed to sample 42 students from each of the 150 schools for a total sample of 6,300 students, while countries/economies taking the PBA needed to sample 35 students from each of the 150 schools for a total sample of 5,250 students. CBA countries/economies taking the financial literacy domain were also required to sample more schools and/or more students per school to obtain an additional sample of 1,650 students, resulting in a total sample of 7,950 students. This group of 1,650 students who took the financial literacy sample was randomly equivalent to, albeit different from, the “main sample” students who did not take financial literacy.

With mathematics as the major domain, one hour of mathematics was administered to most of the students in the main sample (i.e., 96% with CrT and 94% without CrT), and the other domains were only administered to a subset of students.

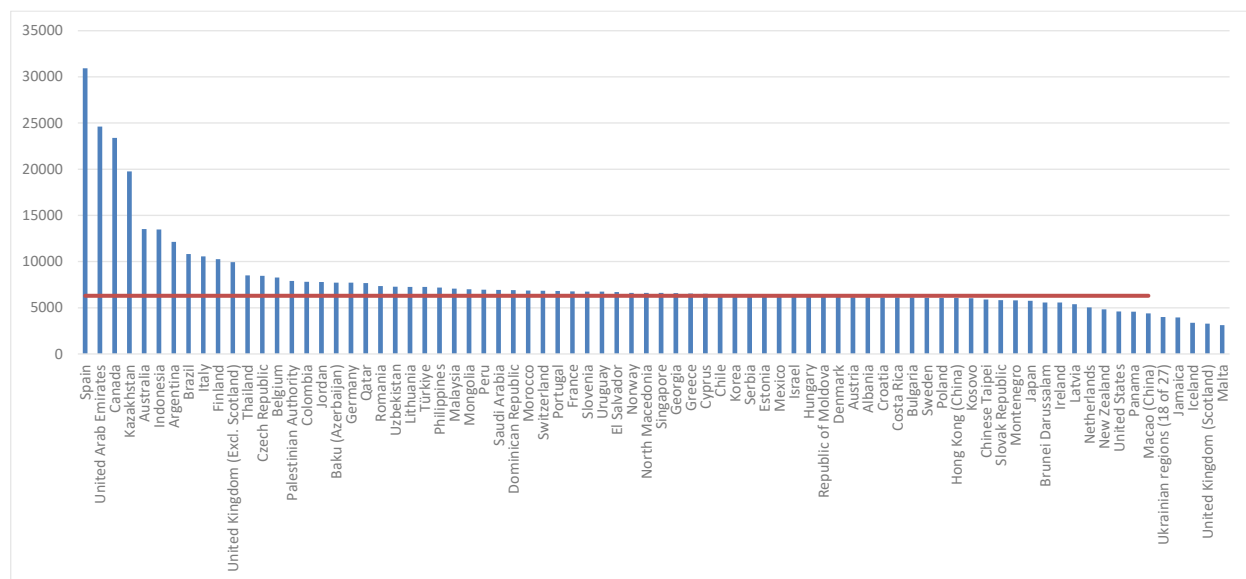
#### *Data yield*

Annex Table 11.A.2 shows the assessment languages and the sample sizes for each of the participating countries/economies. For a student to be considered a “respondent” for PISA, the student needed to meet at least one of the following two criteria: 1) answered more than half of the cognitive items from the

assigned form/booklet, or 2) answered at least one cognitive item and at least one item regarding home possessions (i.e., ST251 or ST255).

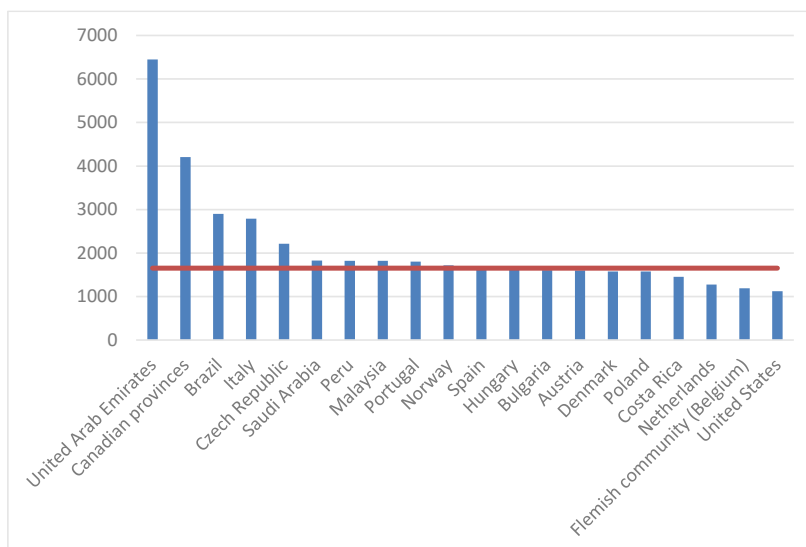
Figure 11.1, Figure 11.2, and Figure 11.3 show the extent to which each country/economy participating in the CBA, the financial literacy assessment, and the PBA met or exceeded the sample size requirements. In each figure, the red horizontal line indicates the sample-size requirements for each design option. Some countries/economies exceeded the requirements because they oversampled certain regions and/or minority languages. As expected, a few countries/economies did not reach the sample size requirements because of their small total population size. Because of on-going post-Covid challenges, 26 countries/economies did not reach their sample-size target. Nevertheless, most of them managed to get very close, and all collected enough data to contribute to the international scaling and to produce high-quality population modelling outcomes that are comparable to those of all other participating countries/economies.

**Figure 11.1. Main sample yield for countries/economies participating in the CBA**



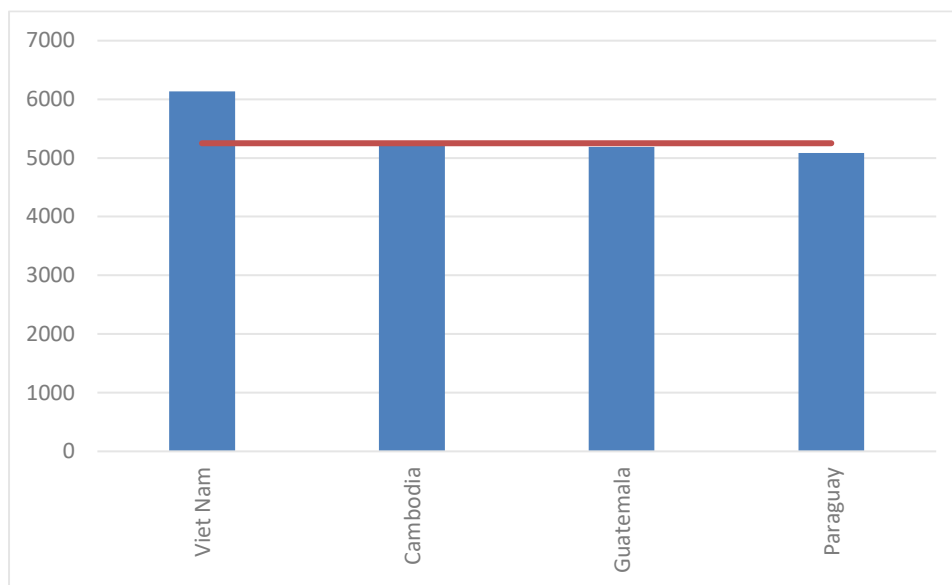
Note: Ukrainian regions (18 out of 27) administered the assessment.

**Figure 11.2. Financial literacy sample yield for participating countries/economies**



Note: 'Canadian provinces' refer to the seven provinces of Canada that participated in the PISA 2022 financial literacy assessment: British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario and Prince Edward Island. It is not a nationally-representative sample. 'Flemish community (Belgium)' refers to the Flemish-speaking population of Belgium. It is not a nationally-representative sample.

**Figure 11.3. Main sample yield for countries/economies participating in the PBA and new PBA**



Since the sample sizes varied greatly across countries/economies, the number of sampled schools and the sample sizes from each school varied as well. As shown in Annex Table 11.A.2, the number of schools ranged from 46 to 983, but most countries/economies met the requirement to sample a minimum of 150 schools.

The PISA 2022 assessment design also required that students be randomly assigned to forms in the prescribed proportions. Results showed that this condition was met for all participating countries/economies and that the assignment of students to items was appropriate for the item analyses and IRT scaling.

### *MSAT data yield for mathematics and reading*

The goal of the mathematics and reading MSAT designs was to improve the measurement precision across a wide range of proficiencies, and at the same time, to collect optimal data needed for the item analyses and IRT scaling. Therefore, it was important to verify that the MSAT design was implemented as intended. Note that some students in some countries/economies took a shorter non-adaptive *Une-heure* (UH) booklet/form. Also, in Israel, some students took a non-adaptive *Ultra-Orthodox* (UO) form. These UH and UO cases were excluded from the MSAT analyses reported in this chapter.

Four critical aspects of the MSAT designs were closely monitored:

- Random assignment to each routing testlet
- Random assignment to each of the adaptive (75%) or linear MSAT paths (25%) in mathematics
- Random assignment to each of the Design A (75%) or Design B (25%) paths and misrouting in the expected proportions, in reading
- Adaptive second and third testlet selection according to students' observed performance on prior stages according the MSAT design.

Near uniform proportions of the total number of alternatives were observed that confirmed the random assignments to the routing testlets, the alternate MSATs (Groups A, B and C in mathematics and Designs A and B in reading<sup>1</sup>), and the mathematics adaptive and linear paths.

The adaptive routing through the mathematics and reading designs are summarized in Figure 11.4a. and Figure 11.4b, showing the proportion of students in each country/economy who were routed to difficult, medium or easy testlet combinations such as: hard testlets in both Stage 1 and Stage 2 in mathematics or reading (HH); or low and medium or hard and medium testlet combinations in mathematics (LM or HM), or low and hard or hard and low difficulty testlets combinations in reading (LH or HL); or low difficulty testlets in both Stage 1 and Stage 2 in mathematics or reading (LL). Students' paths were categorized as missing/undetermined when they did not complete the routing stage or stage 1 and their full path could not be determined by the adaptive algorithm. Note that the 25% of students who were assigned to non-adaptive paths in the hybrid MSAT design are not included in Figure 11.4a.

In both figures, the lowest to highest performing countries/economies are shown from left to right. As intended by design, in the lower-performing countries/economies, a smaller proportion of students were assigned to the most difficult testlets, while in the higher-performing countries/economies, a smaller proportion of students were assigned to the easiest testlets. Also, as intended, every type of testlet was assigned to a high enough proportion of the total sample in each country/economy in each stage, regardless of the proficiency distribution in the country/economy. For reading this was achieved through the misrouting of some students, while for mathematics this was achieved by randomly assigning 25% of students to non-adaptive paths of the hybrid MSAT design. Altogether the observed results confirmed that the MSAT delivery platform worked as intended, and that regardless of the countries/economies' proficiency distributions, the adaptive design always provided the minimum number of responses per item needed for IRT scaling and an appropriate item coverage across the full range of student proficiency.

### **Classical test theory statistics: Item analysis**

Classical item analyses (IA) were conducted on all paper-based and computer-based test items at the national and international levels to verify that the items functioned appropriately. Unexpected results were identified and explored for any indication of possible issues related to data collection, human- or machine-

scoring, or other issues. Descriptive statistics for the observed responses and various missing response codes were provided to countries/economies and the OECD for their review and feedback. Classical item analysis also provided additional descriptive information useful for the review of the IRT modelling outcomes.

The following statistics were computed:

- item response category statistics, including frequency and criterion score mean, standard deviation, and biserial correlation
- (classical) item difficulty
- (classical) item discrimination

Item response categories included several types of non-response and item score categories. An item response was recoded as *not-reached* when a student did not answer the item or any subsequent item in the cluster for non-adaptive domains (science, financial literacy, and creative thinking) or in the MSAT sessions for reading and mathematics. An item response that did not perform properly in the field or had a missing human-coded response code was also converted to not-reached. An item response was recoded as *omitted* when a student did not answer the item but answered one or more of the subsequent items in the cluster or the MSAT path. The category *off-task* was used to identify an invalid missing category when a student did not answer the question in the expected way (e.g., by giving a response not associated with the item or responding with more than one answer in an exclusive choice question). In the computation of the item statistics and in the scaling analyses, the not-reached responses were excluded (i.e., treated as missing/ not-administered), but the omitted and off-task responses were treated as incorrect.

The mean score, standard deviation, biserial/polyserial correlation, and point biserial/polyserial correlation were based on the total block/cluster score where the item appeared.

Statistics for trend items were compared with results from prior PISA cycles. Also, statistics were compiled separately for the PBA and CBA and were examined at the aggregate level across countries/economies. Analyses were also performed separately for each country/economy to identify outlier items that worked poorly or differently across assessment cycles and/or across countries/economies and to detect flaws or obvious scoring rule deviations. Analyses were also conducted by language within each country/economy. UH booklet results were provided for countries/economies, where applicable.

Annex Table 11.A.3 and Annex Table 11.A.4 show examples of the item analysis outputs. Annex Table 11.A.3 shows the IAs of the first three items in block/cluster M01 of one country/economy. The first item, DM033Q01C, is the scored version of the paper-based item PM033Q01 (the corresponding CBA item is CM033Q01), a multiple-choice item. Each section of the table represents one item, and the columns represent the different response categories. The *total* column includes the summary information for all categories, excluding the not-reached (*NOT RCH*) category. The last row (*RSP WT*) shows the scores associated with each response category and the maximum score that can be obtained on the item.

The biserial (*R BIS*) statistic is used to describe the relationship between performance on a single test item and a criterion (usually the total score on the test). It is estimated using the polyserial method which is a generalized form of the correlation between the criterion (which is treated as a continuous variable) and the item score, where the item score is either 0, 1 (for dichotomous items) or 0, 1, 2, 3, ..., *k* (for polytomous items).

The delta statistic is an index of item difficulty based on P+ (proportion correct, or percent correct when expressed as a percentage) which has been transformed so that it is on a scale with a mean of 13.0 and a standard deviation of 4.0. Delta statistics ordinarily range from 6.0 for a very easy item (approximately 95% correct) to 20.0 for a very difficult item (approximately 5% correct), with a delta of 13.0 corresponding to 50% correct.

Annex Table 11.A.4 has two parts. The first part shows a breakdown of the score categories and biserial correlations by category. The second part contains summary data for each item and reveals items that were flagged for surpassing certain thresholds. The thresholds are provided in Annex Table 11.A.5. In this example, the third item is flagged for having an omit rate of greater than 10%.

### **Response time analyses**

The computer-based platform captured response time data for all computer-based items delivered in the CBA countries/economies in both the field trial and main survey. Timing data can be informative in evaluating the level of student engagement and effort over the two-hour testing period. Very little time spent on the assessment was interpreted as low effort, while too much time spent on the assessment (or parts of the assessment) could be an indication of technical problems or low ability. Response time information was aggregated by testlet, cluster, domain, and for the full assessment. Item response times by position and proficiency level were also computed. Overall, results indicate that the CBA data provided valid information that can be used to model items and estimate student performance within and across countries/economies.

#### *Outliers*

Students were generally expected to complete the cognitive assessment within two one-hour periods separated by a break. Within each hour, students followed the prescribed order of clusters or MSAT testlets and units at their own pace. Except for the CBA reading and mathematics assessments, students were expected to complete two 30-minute clusters within an hour, regardless of the positions within the assessment (e.g., clusters 1 and 2 in the first hour, clusters 3 and 4 in the second hour). Within each hour, students were allowed to manage their time between the two assigned clusters. For reading, students were expected to complete the reading fluency items within a 3-minute limit and three self-paced MSAT routing, stage 1 and stage 2 testlets (i.e., testlet 1, 2 and 3) within the remaining time in the hour. For mathematics, students were expected to complete three self-paced MSAT or linear testlets within the hour.

Focusing on larger-than-expected cluster or testlet response times, outliers were identified using the median absolute deviation (MAD) approach (Leys et al., 2013<sup>[1]</sup>; Rousseeuw and Croux, 1993<sup>[2]</sup>). That is, response times greater than  $\text{median}\{x_i\} + 4.4478 * \text{median}\{|x_i - \text{median}(x_j)|\}$ , where  $\{x_i\}$  is the collection of all sample values and  $|\cdot|$  denotes their absolute value, were identified as outliers. Note that in this calculation, median values were identified using international data, not country/economy-level data. This way, the same criterion was used across countries/economies, and the identification of outliers was more stable.

Annex Table 11.A.6 shows the percentages of response time outliers by domain. The proportions of outliers were small—between 0.5% to 1.2% across all domains. Note that, because reading fluency was very short and strictly time-limited, an outlier analysis was not needed.

#### *Cluster- or testlet-level response time*

Annex Table 11.A.7 presents descriptive statistics for testlet or cluster response times for all CBA domains, excluding reading fluency. These values are the sum of the time each student spent on each item in a testlet or cluster, aggregated across students, countries/economies, and positions. Similarly, Annex Table 11.A.8 presents descriptive statistics for domain time, computed as the aggregated item time.

These results show that most students spent a reasonable amount of time on each cluster (with most taking more than 13 minutes and less 30 minutes, approximately from the first (Q1) to the third quartiles (Q3)) or on each testlet (more than 13 minutes and less 30 minutes, approximately Q1 and Q3) or each testlet (more than 6 minutes and less than 22 minutes, approximately Q1 and Q3). However, as sample maximum (MAX) values show that some students did take a large amount of time to complete a given 30-minute cluster, thus and having very little time to finish the subsequent cluster with which it was paired.

Similarly, for mathematics and reading, values show that some students did take a large amount of time to complete the first or the first two testlets and have little time for the subsequent(s) one(s). It is also notable that the last mathematics and reading testlets generally took less time than the other testlets.

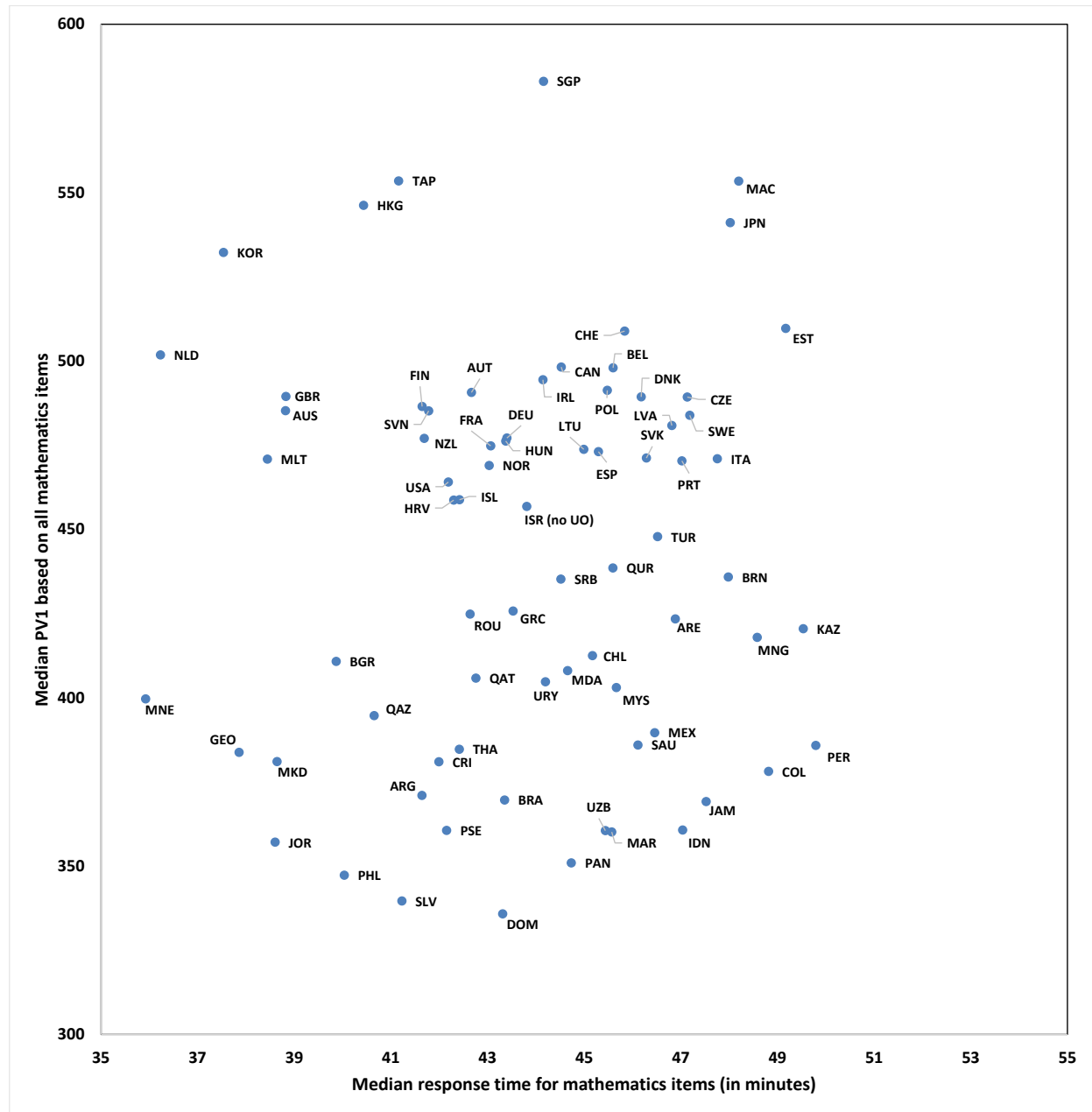
Total domain time was also appropriate in all domains, with most students spending more than 30 minutes (Q1) and less than 54 minutes (Q3). Overall, the time spent in each domain was quite similar, although science and financial has larger Q3 and MAX values. Also, a desired confirmation was that there was no evidence of a timing mode effect between the linear and MSAT groups in mathematics and between design A and B in reading.

### *Response time and student performance*

The relationship between response time and student performance was examined using the median of the cluster-level response time and proficiency levels. The proficiency levels were computed based on the first plausible value (PV1) and a detailed description of their interpretation and cut-offs can be found in Chapter 17. Tables 11.7a – 11.7d show a very similar pattern across all domains and MSAT designs, where from Below Level 1 and up to Level 4, more able students generally spent more time completing each domain. The increase in time spent was most noticeable between students below Level 1 and up to Level 3; then, time spent tapered off up to Level 5 and slightly decreased at Level 6. Again, there was very little difference between the linear and MSAT mathematics tests, except at Levels 5 and 6 where the MSAT students spent about one to two minutes more in median time than the linear students.

While the more proficient students generally took more time to complete the test, median time and median performance varied noticeably across countries/economies. However, as Figure 11.4 shows for mathematics, while countries/economies do vary noticeably in their median PV1 proficiency, there was no clear relationship between median proficiency and median total item response time across at the country/economy level. For example, KOR and SGP, both have high median mathematics scores, but SGP's median response time is close to the overall median response time, while KOR's is well below it.

Figure 11.4. Mathematics median response time by median proficiency across countries/economies

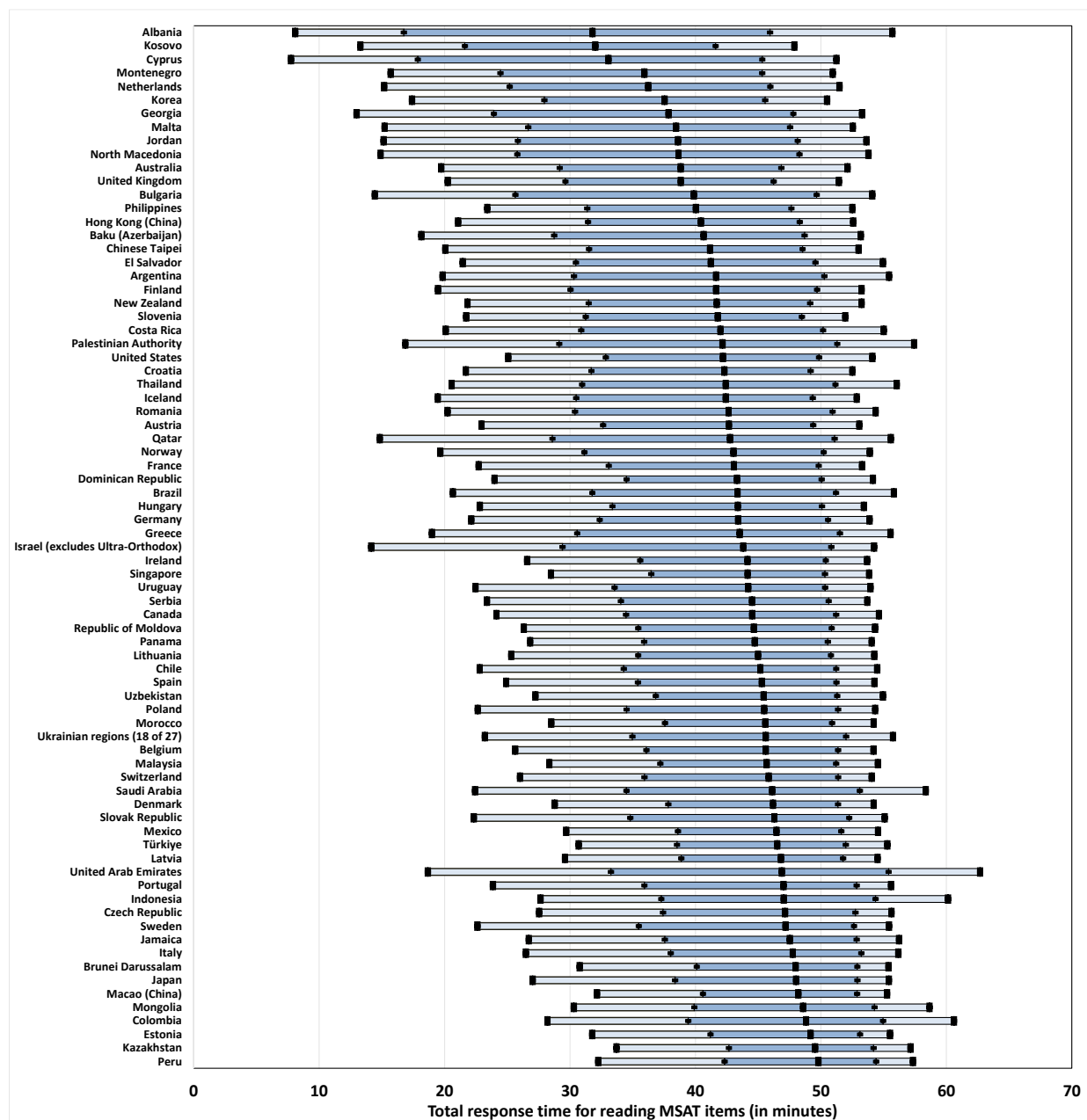


Note: Statistics were calculated only with students who had timing data, excluding UH students. For Israel, Ultra Orthodox students were excluded.

Because of differences in proficiency and other factors, the time it takes students to complete the assessment is expected to vary within each country/economy. This is shown in Figure 11.5 which presents the distribution of the total time spent on the mathematics items for all countries/economies, sorted by the median response time. Note that in a few cases the 90th percentile time was above 60 minutes allocated. This was because the time limit was not strictly enforced to allow for students to finish tasks they were in the middle of.



Figure 11.5. Distribution of mathematics response time in each country/economy



Note: For each country/economy, the solid black line in the middle shows the median total response time, the dark blue horizontal bars range from the 25th to the 75th percentiles, and the light blue horizontal bars range from the 10th to the 90th percentile. Countries/economies are sorted by their median MSAT response time. For Israel, Ultra Orthodox students were excluded.

### Item-level response time

Response time and the relationship between response time and performance were also explored at the item level.

Figure 11.7 and Figure 11.8. show the median item-level response time (aggregated across all countries/economies) for the trend and new mathematics items, respectively, disaggregated by students' proficiency levels based on PV1. For most but not all items, as we have seen above with the total domain

time, low-performing students (blue and red lines) had similar and relatively short response times, while high-performing students (green and purple lines) had longer response times and larger variability in the response times. This pattern was consistently observed for both the trend and new mathematics items. Furthermore, there are some clear peaks indicating items on which high-performing students spend substantial more time than low-performing students.

For the creative thinking items, median item response times are shown in Figure 11.9, for each country/economy. A similar approach was employed but levels were calculated using the first non-linear score transformed value instead of PV1. More detail on the CrT scores and their levels can be found in Chapter 18. As expected, since items were typically more demanding and fewer of them were administered, students generally spent more time per item than for the other domains. Across countries/economies, the amount of time spent per item varied, however, the timing patterns across items were similar.

### ***Response time reflecting possible motivation or administration issues***

On average, students completed the entire test in 83.34 minutes (excluding a short break between the two assessment hours), with a standard deviation of 21.74 and a median of 87.46 minutes. Some students completed the test in less than 30 minutes (found in all countries/economies, 2.7% of the overall sample), while some students took longer than 120 minutes to complete the test (1.5% of the overall sample). At the country/economy-level, students in Kazakhstan, Peru, Mongolia, and Macao took the longest time to complete the entire test, with a median time of 100.8, 99.0, 99.0 and 98.9 minutes, respectively. Students in Cyprus, Albania, and Kosovo took the shortest time to complete the test, with a median time of 67.2, 67.7 and 69.5 minutes, respectively.

There were five countries/economies where 5% or more of the students exceeded the time limit: United Arab Emirates (11.9%), Indonesia (9.9%), Colombia (7.6%), Mongolia (5.7%) and Saudi Arabia (5.4%). This could be explained by students in these countries generally spending more time to complete the test and by the fact that time limits were not strictly enforced so that students in the middle of a task could finish without being abruptly cut-off. Apart from these countries/economies, only a small proportion of respondents in each country/economy had very long or short total response times, indicating that there were no systematic administration and/or motivation issues. Furthermore, students with these extreme response times appeared to be randomly distributed across schools and countries/economies.

### ***Position effects***

According to the PISA test design, each student takes one of many alternative test forms made up of different clusters/testlets in different positions. For example, a student may take two science clusters in the first hour and then take three mathematics testlets in the second hour, while another student may take the same domains, but in the reverse order. Item position effects are a concern in large-scale assessments because substantial position effects, if present, would increase measurement error and may introduce bias in parameter estimation. To mitigate any potential item position effects, as in previous cycles, the PISA 2022 main survey design balanced the order of the domains (between the first and second hour) as well as the order of the clusters or testlets within each domain (see Figure 2.5 in Chapter 2 for the full form design used in PISA 2022). Thus, PBA and CBA clusters and items within them (in fixed position) appeared in the first hour in positions 1 and 2 and in the second hour in positions 3 and 4. The CBA testlets for mathematics appeared in the first hour in positions 1, 2, and 3, and in the second hour in positions 4, 5, and 6. The exception was reading, where the MSAT design was partially balanced with the core testlets appearing in positions 1 and 4 and the stage 1 and stage 2 testlets each appearing in positions 2, 3, 5, and 6.

As prior PISA cycle results have indicated, the PISA 2022 results summarized below show that position effects are significant and justify the use of the complex BIB and balanced MSAT designs implemented to minimize their impact.

To evaluate and confirm that the impact of item positions studied in the field trial was minimal in the PISA 2022 main survey, position effects were examined in terms of: 1) proportion of correct responses, 2) median response time, and 3) rate of omitted responses. For PBA and CBA domains, cluster-level statistics are reported for positions 1, 2, 3, and 4, and position effects are reported as the difference between positions 4 and 1. For the mathematics and reading MSATs, domain-level statistics were reported for the 1<sup>st</sup> hour and the 2<sup>nd</sup> hour<sup>2</sup> and the position effects are reported by the difference between hour 2 and hour 1.

Annex Table 11.A.9. Median response time (in minutes) by cluster position in the CBA for non-adaptive domains and Annex Table 11.A.10 present the position effects in terms of the median response time<sup>3</sup> averaged by cluster position and by assessment hour, respectively. For all domains, students spent more time on a cluster when presented in position 1 than in position 4. Financial literacy items had a noticeably higher median response time when in cluster position 1, resulting in a larger difference between the median response times for cluster positions 1 and 4. There were indications that some students spent much more time on clusters 1 and 3, leaving them with less time for clusters 2 and 4, respectively. Annex Table 11.A.9 shows that the position effects by hour were generally smaller than the position effects by cluster. Across domains, students spent between 3.54 to 6.92 minutes less in median response time in the second hour. For mathematics, positions effects appear nearly identical between the linear and MSAT part of the hybrid design. For reading, the response-time position effect is larger for the core than the first and second stages.

Annex Table 11.A.11 and Annex Table 11.A.12 present the position effects in terms of the average P+, averaged by cluster position and by assessment hour, respectively. By cluster, the decreases in P+ between position 1 and 4 ranged from 0.051 in creative thinking to 0.89 in financial literacy. Overall, cluster position effects were similar to values observed in prior PISA cycles. By assessment hour (Annex Table 11.A.9), for all non-adaptive domains, a smaller decrease in P+ between the 1<sup>st</sup> and 2<sup>nd</sup> assessment hour was observed compared to the decrease in P+ between the 1<sup>st</sup> and 4<sup>th</sup> cluster position. For the mathematics linear and adaptive MSAT trend and new items, the decrease in average P+ between the 1<sup>st</sup> and 2<sup>nd</sup> hour were all relatively small and similar to the decreases observed in the other domains.

The proportions of omitted responses at different positions for all CBA countries/economies were analysed to further examine the quality of data affected by position. The proportion of omitted responses are shown by cluster position and assessment hour in Annex Table 11.A.13 and

Annex Table 11.A.14, respectively. These do not include the 'not-reached' items. Note that the proportion of omitted responses for reading fluency are 0 because students had to respond to each item presented (i.e., they were not able to skip the item). Overall, the omission rates by cluster and by hour were very similar across the domains. As in PISA 2018, the omission rates for all domains in all positions were less than 0.10, and the omission rates in positions 2 and 4 were higher than the rates in positions 1 and 3, respectively.

Position effects were also reviewed for the new PBA forms. Annex Table 11.A.15 and Annex Table 11.A.16 report the average P+ and the average omission rates by cluster position. By comparison with the results from the PBA forms used in the prior cycles, the new PBA position effect were noticeably smaller: Position 4 – Position 1 decrease in P+ by less than 0.04 (compared to less than 0.09) and Position 4 – Position 1 omits increased by less than 0.02 (compared to less than 0.05).

## IRT modelling and scaling

The modelling and scaling of the PISA 2022 main survey data followed the general approach developed for PISA 2015 [OECD (2017<sub>[3]</sub>), Chapter 9]. The following sections describe the IRT models and their assumptions, as well as the IRT scaling approach used in PISA 2022. The scaling issues associated with the mathematics and reading MSAT designs and how they were resolved are addressed as well.

### IRT models and assumptions

As in PISA 2015 and 2018, the unidimensional multiple-group IRT model (Bock and Zimowski, 1997<sub>[4]</sub>; von Davier and Yamamoto, 2004<sub>[5]</sub>) based on the two-parameter logistic model (2PLM) (Birnbaum, 1968<sub>[6]</sub>) for the binary item responses and the generalized partial credit model (GPCM) (Muraki, 1992<sub>[7]</sub>) for the polytomous item responses were used for each domain. The 2PLM is a generalization of the Rasch model (Rasch, 1960<sub>[8]</sub>), which assumes that the probability of a correct response to item  $i$  depends only on the difference between the student  $v$ 's trait level  $\theta_v$  and the difficulty of the item  $b_i$ . In addition, the 2PLM postulates that for every item, the association between this difference and the response probability depends on an additional item discrimination parameter  $a_i$ :

$$P(x_{vi} = 1 | \theta_v, b_i, a_i) = \frac{\exp(Da_i(\theta_v - b_i))}{1 + \exp(Da_i(\theta_v - b_i))} \quad \text{Formula 11.1}$$

The probability of a positive response (e.g., solving an item correctly) is strictly monotonic, increasing with  $\theta_v$ . The item discrimination parameter  $a_i$ , usually scaled by a constant  $D = 1.7$ , characterizes how quickly the probability of solving the item approaches 1.00 with increasing trait level  $\theta_v$ , when compared to other items. In other words, the model accounts for the possibility that responses to different items do not have the same weight with relation to the latent trait. The discrimination parameter  $a_i$  describes how well a certain item relates to the latent trait and, therefore, discriminates between examinees with different trait levels compared to other items on the test. One important special case of the model is when  $a_i = 1$  for all items, in which case, the model is equivalent to a Rasch model.

The GPCM (Muraki, 1992<sub>[7]</sub>), like the 2PLM, is a mathematical model for the probability that an individual will respond in a certain response category on a particular item. While the 2PLM is suitable for items with only two response categories (dichotomous items), the GPCM can be used with items with more than two response categories (polytomous items). The GPCM reduces to the 2PLM when applied to dichotomous responses. For an item  $i$  with  $m_i + 1$  ordered categories, the probability of obtaining a score of  $k$  ( $0, 1, 2, \dots, m_i$ ) under the GPCM can be written as:

$$P(x_{vi} = k | \theta_v, b_i, a_i, d_i) = \frac{\exp\{\sum_{r=0}^k Da_i(\theta_v - b_i + d_{ir})\}}{\sum_{u=0}^{m_i} \exp\{\sum_{r=0}^u Da_i(\theta_v - b_i + d_{ir})\}} \quad \text{Formula 11.2}$$

where  $d_{ir}$  is the item-category threshold or step parameter as indicated in Appendix A), with  $\sum_{r=1}^{m_i} d_{ir} = 0$  and  $d_{i0} = 0$ .<sup>4</sup>

Critical assumptions of most IRT models and the models used in PISA are conditional independence (sometimes referred to as local independence) and unidimensionality. Under conditional independence, item response probabilities depend only on the latent trait and the specified item parameters—there is no additional dependence on any demographic characteristics of the students, responses to any other items presented in a test, or the survey administration conditions. Under the unidimensionality assumption, a common single latent variable accounts for performance on the full set of items. With past PISA data, these assumptions have been verified and item parameters have been estimated for each cognitive domain

separately through unidimensional IRT models. These assumptions need to be confirmed for each domain in which any new items are used.

With these assumptions, we can formulate the following joint probability of a particular response pattern  $x_v = (x_{v1}, \dots, x_{vn})$  across a set of  $n$  items:

$$P(x_v | \theta_v, \boldsymbol{\beta}) = \prod_{i=1}^n P(x_{vi} | \theta_v, \boldsymbol{\beta}_i), \quad \text{Formula 11.3}$$

where  $\boldsymbol{\beta}_i$  is the vector of parameters for item  $i$  from the associated IRT model. 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 (indexed  $v=1, 2, \dots, N$ ) provide their answers independently of one another and that the student's proficiencies are sampled from a distribution  $f(\theta)$ . Using the sampling weights  $w_v$ , the likelihood function is, therefore, characterised as:

$$P(\mathbf{X} | \boldsymbol{\beta}) = \prod_{v=1}^N w_v \int P(x_v | \theta, \boldsymbol{\beta}) f(\theta) d\theta. \quad \text{Formula 11.4}$$

Typically, the item parameters that provide the best possible fit to a given data set are estimated by maximising this function through a process called *item calibration*. The item parameters can then be used in the subsequent analyses, such as in the estimation of individual plausible values and population characteristics. However, it should be noted that IRT modelling does not provide an absolute scale, since any linear transformation of the item and latent trait parameters in the above formula leads to the exact same likelihood function, often referred to as scale indeterminacy or non-identifiability. Therefore, as part of the calibration process, a choice must be made for the IRT scale to be determined.

For further information regarding the IRT models discussed, see Fischer and Molenaar (1995<sub>[9]</sub>), van der Linden and Hambleton (1997<sub>[10]</sub>; 2016<sub>[11]</sub>), or von Davier and Sinharay (2014<sub>[12]</sub>) for the use of these models in the context of international comparative assessments.

### **IRT item calibration and scaling**

The PISA data collection designs are complex, and the assessments are adapted and translated for each participating country/economy into one or more languages. To better account for potential cultural and language differences, and to optimally scale the item parameters and proficiency estimates across countries/economies and across modes (PBA and CBA), new calibration and scaling approaches were implemented in 2015. For each domain, a series of multi-group concurrent calibrations of the historical data (2015 and prior PISA cycles) were conducted (von Davier et al., 2019<sub>[13]</sub>) (OECD, 2017<sub>[3]</sub>), Chapter 9. As a result, all the items used in all the PISA cycles up to 2015 were estimated and scaled onto new common IRT scales (by domain) and new transformations from these IRT scales to the existing PISA reporting scale were established to preserved trend comparability.

For the first run of the series of multi-group concurrent calibrations, the item parameters were constrained so that only one set of *common or international parameters* was estimated per item to model the data for all the country-by-language-by-cycle groups. As part of the calibration process, the fit of the common item parameters to the data for each pre-defined group was evaluated. Then, item-by-group interactions were identified when the fit to the data was found to be poor (i.e., the value of the item fit statistic, discussed below, was higher than a chosen threshold value). In the subsequent runs, new *unique or group-specific*

item parameters were estimated in the group or groups in which misfit was found and the item fit threshold was gradually lowered until the ultimate target threshold was reached, thus allowing additional group-specific item parameters to be estimated. The fundamental consideration of using this stepwise procedure is to optimize both the model data fit and the comparability across all groups—keeping common item parameters for as many groups as possible or minimizing the use of unique parameters. By allowing unique item parameters for items that show item-by-group interactions – in contrast to excluding such items or accepting poor common item parameter fit – the measurement error is reduced without introducing bias. The research base for this approach can be found in Meredith (1993<sub>[14]</sub>); Reise, Widaman and Pugh (1993<sub>[15]</sub>); Glas and Verhelst (1995<sub>[16]</sub>); Yamamoto (1997<sub>[17]</sub>); Glas and Jehangir (2014<sub>[18]</sub>); Meredith and Teresi (2006<sub>[19]</sub>); as well as Oliveri and von Davier (2011<sub>[20]</sub>; 2014<sub>[21]</sub>).

Since PISA 2015, in 2018 and now in 2022, the same IRT calibration and scaling approach has been used to estimate new item parameters onto the existing IRT scales. However, the historical data no longer needed to be included in the scaling since all trend items (reused from 2015 and/or prior PISA cycles) had already been calibrated and scaled. Therefore, in PISA 2022, as in PISA 2018, a fixed item parameter linking approach was utilized with the trend item parameters fixed to their values established in the 2015 and 2018 scaling in the first calibration run to start the estimation of international parameters for the new items. The subsequent runs, then proceeded in the same manner as described above to evaluate item-by-country-by-language interactions (i.e., group-level item-fit) and to estimate unique parameters when needed.

Group-level item-fit analyses are a critical part of the scaling analyses described above. Different types of differential item functioning (DIF) statistics can be used to evaluate the extent to which the IRT model applied to a group fits the response data collected from that group. In the context of the IRT models used in since PISA 2015, the extent to which the model-based item characteristic curve (ICC, computed using formula 11.1 or 11.2 for the 2PLM or the GPCM) and the empirical ICC can differ is evaluated based on the mean deviation (MD) and the root mean square deviation (RMSD) statistics:

$$MD_g = \int [p_g^{obs}(\theta) - p_g^{exp}(\theta)] f_g(\theta) d\theta,$$

Formula 11.5

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

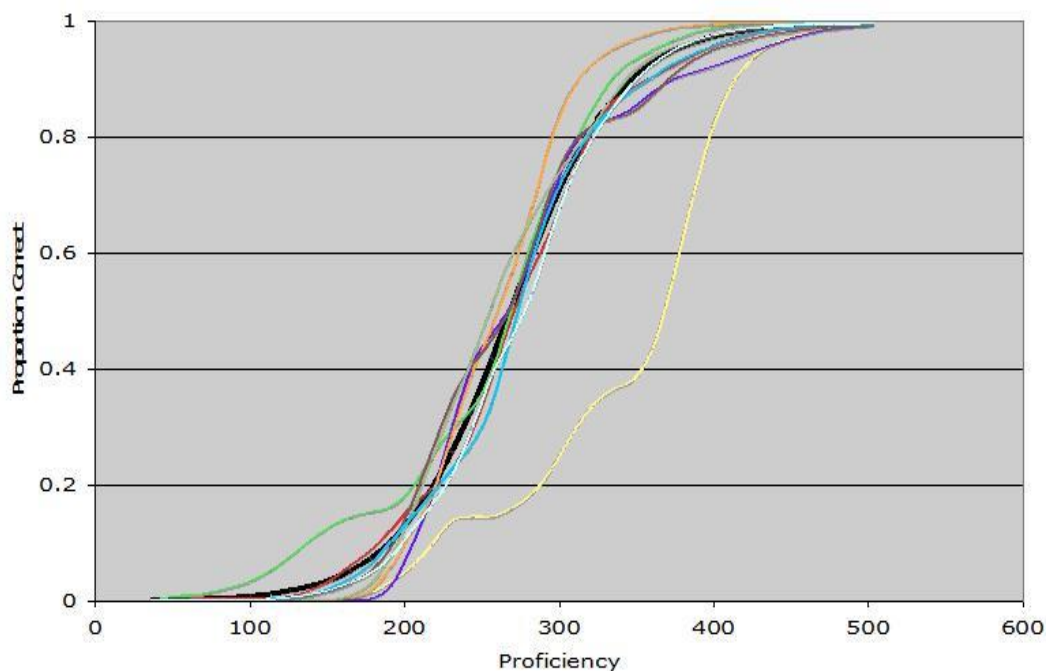
Formula 11.6

where  $g = 1, \dots, G$  is a country-by-language group;  $p_g^{obs}(\theta)$  and  $p_g^{exp}(\theta)$  are the observed and expected probability of a correct response given proficiency  $\theta$ ; and  $f_g(\theta)$  is the group-specific density on the students' ability scale (Khorramdel, Shin and von Davier, 2019<sub>[22]</sub>; von Davier, 2005<sub>[23]</sub>). The observed probability correct is based on the pseudo counts from the expectation-maximization (EM) algorithm that is used to estimate the model (Bock and Aitkin, 1981<sub>[24]</sub>), while the expected probability correct is based on the estimated item parameters. The moments of the group-specific densities are also estimated for each country-by-language group (Xu and von Davier, 2008<sub>[25]</sub>).

The observed item characteristic curve (ICC) is obtained from the observed responses across students for each item, and the expected ICCs are computed based on the IRT model using the estimated item parameters. RMSD quantifies the magnitude and MD quantifies the magnitude and direction of deviations in the observed data from the estimated common or group-specific item characteristic curves for each single item. However, while MD is sensitive to the difference in observed and model-based item difficulty represented by the  $b$  parameter in formulae 11.1 and 11.2, RMSD is sensitive to the differences in both item difficulty and item discrimination represented by the  $a$  (or slope) parameter in formulae 11.1 and 11.2.

To demonstrate the use of item fit statistics (RMSD, MD), Figure 11.6 shows one example plot for a dichotomously scored item estimated via the 2PLM. It illustrates how the common item parameter fits data from all groups, except for one group. In the figure, the solid black curve is the model-based 2PLM item response curve that corresponds to the common item parameters; the other lines are observed proportions of correct responses along the proficiency scale (horizontal axis) for the data from each group. This plot indicates that the IRT model-based curve conforms to the observed data; proportions of correct responses given the proficiency are quite similar for most countries/economies. However, the data for one country/economy, indicated by the yellow line, shows a noticeable departure from the common item characteristic curve and curves for other groups. This item is far more difficult in that particular country/economy, conditional on proficiency level. Thus, a unique set of parameters would be estimated for this item, for this group.

**Figure 11.6. Item response curve (ICC) for an item where the common item parameter is not appropriate for one group**



### ***Calibration and scaling of the mathematics and reading adaptive domains***

The purpose of adaptive testing is to better match test difficulty with student proficiency and avoid administering items that are either too easy or too difficult. Unlike data collected using traditional linear testing, this results in some of the data (responses to some of the relatively easy or difficult items) being missing not at random and a reduced overlap between test forms delivered to students having different proficiency levels. Unfortunately, using such data for IRT scaling could lead to bias in the item parameters and the student proficiency estimates (Jewsbury and van Rijn, 2020<sup>[26]</sup>). To address this issue, many testing programs use a two-step data collection design that allows for item parameters to be pre-calibrated through a non-adaptive data collection. Then, once their item parameters have been established, they are incorporated into the operational instrument administration (Glas, 2010<sup>[27]</sup>). However, for PISA, such approach would require the collection of much larger, population representative, field trial data.

Instead, the PISA reading and mathematics MSATs were designed to ensure both adaptation for many countries/economies performing across wide proficiency ranges, and appropriate data collection for the accurate scaling and estimation of international and unique parameters for all countries/economies. To do so, three issues that could threaten the quality of the reading and mathematics PISA scaling were addressed.

First, in designing and finalizing the MSAT, units were assigned to ensure the linkage across different MSAT forms (i.e., routing paths) through common units appearing multiple times across testlets. Similar to the BIB designs used in earlier PISA cycles, in which the same cluster appears across different forms, such linkage through common units across different testlets was expected to improve the efficiency of the item calibration. Such design considerations were tested and verified with simulation studies before the main survey implementation. Second, a proportion of students were assigned in a non-adaptive manner by overwriting some routing decisions as part of the reading MSAT design or by developing a non-adaptive MSAT assigned to a proportion of students as part of the mathematics *hybrid* MSAT design. In both cases, this ensured that more than 250 responses across the full proficiency range were collected for all items in all countries/economies. Third, the order of position of units within testlets has to vary to be able to adapt and assemble easier and more difficult testlets. See Chapter 2 for more detailed descriptions of the design implemented.

The effectiveness of the PISA MSAT designs was investigated during their development using data simulation and field trial data, and the quality of the designs implemented was confirmed using main survey data.

Within-testlet unit order effects were examined in the 2022 mathematics and the 2018 reading field trials to confirm the invariance of item parameters by unit order (Yamamoto et al., forthcoming<sup>[28]</sup>). If the unit order had shown to significantly impact item parameter and proficiency estimates, an MSAT design could not have been implemented because a significant lack of invariance would undermine the effectiveness of the design. The field trial results confirmed the feasibility of introducing an MSAT into the main survey, as unit order effects were found to be negligible.

Model data fit from the same calibration approach used for other non-adaptive domains and alternatives that incorporated MSAT-specific information, such as routing outcomes to define the group in the multi-group calibration process, were evaluated through simulation studies (van Rijn and Shin, 2019<sup>[29]</sup>). Results showed that incorporating MSAT-specific information in the group definition for the multiple-group IRT model resulted in larger errors in the item parameter estimation. Because routing decisions in PISA are largely based on cognitive responses (i.e., sum scores based on the machine-scored items), using this information again to define groups for the multiple-group IRT model would violate the conditional independence assumptions. In the end, after reviewing the results from calibrating simulated data and the collected main survey data, it was determined that the same approach used for the calibration of the other non-adaptive domains was appropriate. A recent study (Jewsbury et al., 2023<sup>[30]</sup>) also provides theoretical justification for this choice.

### ***Calibration and scaling of reading fluency***

As discussed in Chapter 3, reading fluency items were included as a part of the reading scale, which was assessed principally through the reading MSAT. These items were introduced in 2018 to increase the measurement precision at lower levels of the reading scale. However, as their content and format tend to differ from that of the “regular” reading items, the reading fluency items could affect the existing reading scale. Therefore, following the procedure established in 2018 data (OECD, 2022<sup>[31]</sup>, Chapters 9 and 12), to maintain the existing reading scale and avoid any potential issues that could weaken the comparability of the reading scale across cycles, the calibration of reading fluency items was done after the estimation of reading items had been finalized. That is, after the scaling of “regular” reading items was finalized, the



reading fluency data was added to the reading data and the reading fluency items were scaled. Because all items were trend, their parameters were fixed to their final 2018 values.

## Population modelling and multiple imputation

This section describes the population modelling approach that is employed in the analyses of PISA data that combines the latent regression model for a large number of background variables with the IRT model for cognitive item responses. It also explains the imputation methodology for obtaining plausible values for proficiency (both scales and subscales) and for using these to estimate descriptive statistics for populations and subpopulations. This methodology provides countries/economies with databases that can be used for secondary analyses of relationships between proficiency and background variables.

The prime goal of PISA is to compare the skills and knowledge of 15-year-old students across countries/economies and over cycles, reporting on group-level scores in the core domains of mathematics, reading, and science, as well as other domains (Kirsch et al., 2013<sup>[32]</sup>). For group-level reporting assessments such as PISA, where the number of items that can be administered to each student is limited and where the focus of the assessment is on population characteristics, the use of point estimates could lead to seriously biased estimates of population characteristics (Mislevy, 1991<sup>[33]</sup>; Thomas, 2002<sup>[34]</sup>; von Davier, Gonzalez and Mislevy, 2009<sup>[35]</sup>; von Davier et al., 2006<sup>[36]</sup>; Wingersky, Kaplan and Beaton, 1987<sup>[37]</sup>).<sup>5</sup> Reporting outcomes are not intended to have consequences of any sort for individual students, and test forms are kept relatively short to minimise the testing burden on students. At the same time, PISA aims to provide a broad content coverage of each of the domains through a large number of items organised into different, but linked, test forms. Thus, each student receives a relatively small number of items from two domains in a two-hour testing period.

Population modelling for PISA 2022 followed the same general approach used in previous cycles. This approach incorporates the IRT scaling of the students' cognitive data from multiple domains, and the students' background data specified as covariates (e.g., gender, country/economy of birth, academic and non-academic activities, attitudes, etc.) through multivariate latent regression models (von Davier et al., 2006<sup>[36]</sup>). Data from multiple cognitive domains are modelled together to increase the accuracy of the population estimates in each domain by borrowing information from the other cognitive domains. The *plausible value methodology* uses the latent regression models estimated from each country/economy data to impute multiple proficiency values (plausible values) for each student instead of a single point estimate in each domain. The imputation draws the plausible values from the posterior distributions constructed through the multivariate latent regression model and the student data. The multiple imputations from the posterior distributions can then be used to appropriately account for measurement errors in the relations between (sub)population proficiency distributions and characteristics in the background data.

IRT scaling, latent regression, and multiple imputation are carried out through the following steps:

1. IRT scaling: estimates the item parameters for each domain to provide comparable scales across countries/economies and cycles using the unidimensional IRT models described in Formula 11.1 and Formula 11.2 (see also section “IRT calibration and scaling”).
2. Latent regression: estimates the regression coefficients ( $\Gamma$ ) and the residual variance-covariance matrix ( $\Sigma$ ) using the estimated item parameters from step 1 as true values (Thomas, 1993<sup>[38]</sup>).
3. Multiple imputation: draws ten plausible values for each student on each domain from posterior distributions of proficiency using estimated  $\Gamma$  and  $\Sigma$  (Mislevy and Sheehan, 1987<sup>[39]</sup>; von Davier, Gonzalez and Mislevy, 2009<sup>[35]</sup>).

Because of the large number of background collected, a “divide-and-conquer” approach (Patz and Junker, 1999<sup>[40]</sup>) is used to reduce the computational burden of Step 2 (latent regression) and to avoid over-parametrisation. First, all variables in the BQ are contrast coded.<sup>6</sup> Contrast coding allows for the inclusion

of missing responses and avoids the necessity of assuming a linear relationship between the responses to any question and the outcome variable. Second, a principal components analysis (PCA) is conducted to 1) remove collinearity among variables when present and 2) reduce the large number of contrast-coded BQ variables into a smaller number of principal components that are sufficient to account for a large proportion of the variation in the BQ variables without over-parameterisation. This process is conducted country/economy by country/economy to accommodate common BQ variables collected across all countries/economies, to accommodate optional specific BQ variables of participating country/economy's interest, and to allow for the estimation of country/economy-specific relationships between the BQ data and the proficiency variables.

The country/economy-specific multivariate latent regression gives an expression for student's proficiency distributions on the multidimensional scales conditional on covariates ( $\mathbf{y}$ ) in addition to the item responses ( $\mathbf{x}$ ). Based on Bayes' theorem, the posterior distribution of skills given the observed item responses and covariates (i.e., contextual information) is constructed as follows:

$$P(\boldsymbol{\theta}_v | \mathbf{x}_v, \mathbf{y}_v, \Gamma, \Sigma) \propto P(\mathbf{x}_v | \boldsymbol{\theta}_v, \mathbf{y}_v, \Gamma, \Sigma) P(\boldsymbol{\theta}_v | \mathbf{y}_v, \Gamma, \Sigma) = P(\mathbf{x}_v | \boldsymbol{\theta}_v) P(\boldsymbol{\theta}_v | \mathbf{y}_v, \Gamma, \Sigma), \quad \text{Formula 11.7}$$

where  $\boldsymbol{\theta}_v$  is a vector of length  $D$  with scale values (these values correspond to performance on each of the skills) for student  $v$ . As shown, the posterior distribution of proficiency is proportional to the likelihoods of the item-response data and prior distributions. Given the conditional independence assumption,  $P(\mathbf{x}_v | \boldsymbol{\theta}_v)$  is the product of independent likelihoods for the observed response to each cognitive item (estimated by IRT models) within each scale (i.e., the likelihood is factored). Next,  $P(\boldsymbol{\theta}_v | \mathbf{y}_v, \Gamma, \Sigma)$ , which is a prior distribution, is the multivariate joint density of proficiencies of the scales, conditional on the extracted principal components derived from background responses, and parameters  $\Gamma$  and  $\Sigma$ . Note that Formula 11.7 technically also depends on the item parameters, but these are treated as fixed in the computations in steps 2 and 3 and therefore dropped from the equation.

More precisely, the latent proficiency variables for each student  $v$  are assumed to follow multivariate normal distributions:

$$\boldsymbol{\theta}_v \sim ND(\Gamma' \mathbf{y}_v, \Sigma), \quad \text{Formula 11.8}$$

where  $\Gamma$  is the  $K \times D$  matrix of regression coefficients,  $K$  is the number of conditioning variables (the number of principal components plus a dummy for the intercept), and  $\Sigma$  is the  $D \times D$  residual variance-covariance matrix. As noted, the parameters  $\Gamma$  and  $\Sigma$  are estimated using the estimated item parameters from the first step. Let  $\phi(\boldsymbol{\theta}_v | \Gamma' \mathbf{y}_v, \Sigma)$  denote the multivariate normal density with mean  $\Gamma' \mathbf{y}_v$  and covariance matrix  $\Sigma$ .

Operationally, the procedure is repeated several times to model the main and financial literacy datasets from each country/economy. Once focusing on the core domain data (mathematics, reading, and science; then  $D = 4$ ). Twice focusing on each of the two sets of 4 mathematics subscales data with the reading and science data ( $D = 6$ ). Once focusing on the creative thinking data with the core domains data ( $D = 5$ ). And once focusing on financial literacy with mathematics and reading data ( $D = 3$ ). Latent correlations among those domains are estimated as part of the  $D \times D$  residual variance-covariance matrix.

Involving all students in the country/economy, the weighted likelihood function becomes

$$L(\Gamma, \Sigma; \mathbf{X}, \mathbf{Y}) = \prod_{v=1}^N w_v \int \prod_{d=1}^D P(x_{vd} | \theta_d) \phi(\boldsymbol{\theta} | \Gamma' \mathbf{y}_v, \Sigma) d\boldsymbol{\theta}, \quad \text{Formula 11.9}$$

where  $x_{vd}$  is the vector of item responses of students for dimension  $d$ . As noted above, the item parameters  $\beta_d$  associated with  $P(x_{vd}|\theta_d)$  for dimensions  $d=1, \dots, D$  are estimated in the IRT item calibration stage, prior to the estimation of the latent regression  $\phi(\theta|\Gamma'y_v, \Sigma)$ , and treated as fixed. That is, the latent regression parameters  $\Gamma$  and  $\Sigma$  are estimated conditionally on the previously estimated item parameters  $\beta$ .

As suggested by Mislevy et al. (1992<sup>[41]</sup>), the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977<sup>[42]</sup>) is used for maximizing the likelihood function in Formula 11.9 with respect to  $\Gamma$  and  $\Sigma$ . A multivariate variant of the latent regression model based on the Laplace approximation (Thomas, 1993<sup>[38]</sup>) is applied in reporting PISA proficiencies on more than two dimensions (domains and subdomains).

After the estimation of regression parameters through the EM algorithm is completed, multiple imputations (plausible values) for each student  $v$  are drawn from a normal approximation of the conditional posterior distribution of proficiency. More specifically, plausible values are drawn following a three-step process. First, a value for  $\Gamma$  is drawn from  $N_D(\hat{\Gamma}, \widehat{V}(\Gamma))$  where  $\widehat{V}(\Gamma)$  is the estimated variance of the maximum likelihood estimate  $\hat{\Gamma}$  obtained from the EM algorithm (Rubin, 1987<sup>[43]</sup>). Second, conditional on the generated value for  $\Gamma$  and the fixed value of  $\Sigma = \hat{\Sigma}$  obtained from the EM algorithm, the Laplace approximations to the individual posterior mean and variance are computed denoted by  $\tilde{\theta}_v$  and  $\tilde{\Sigma}_v$ , respectively. In the third step, the  $\theta_v$  are drawn independently from a multivariate normal distribution  $N(\tilde{\theta}_v, \tilde{\Sigma}_v)$  for each student  $v$  (Chang and Stout, 1993<sup>[44]</sup>). These three steps are repeated 10 times, effectively resulting in 10 plausible values for  $\theta_v$  for each student.

## Analysis of data with plausible values

If the multivariate latent proficiencies  $\theta_v$  were known for all students, it would be possible to directly compute any statistic  $t(\theta, y)$ , for example, subpopulation sample means, sample percentiles, or sample regression coefficients, to estimate a corresponding population quantity  $T$ . However,  $\theta$  values are not observed, but estimated latent variables through measurement models. To overcome this problem, the approach developed by Rubin (1987<sup>[43]</sup>) is taken in which  $\theta$  is treated as missing data.

Therefore, the value  $t(\theta, y)$  is approximated by its expectation given the observed data,  $(x, y)$ , as follows:

$$t^*(x, y) = E[t(\theta, y)|x, y] = \int t(\theta, y)p(\theta|x, y)d\theta.$$

Formula 11.10

It is possible to approximate  $t^*$  using plausible values (also referred to as multiple imputations) instead of the unobserved  $\theta$  values. A replication approach [see, e.g., Johnson, (1989<sup>[45]</sup>); Johnson and Rust (1992<sup>[46]</sup>); Rust, (2014<sup>[47]</sup>)] is used to obtain a variance estimate for the proficiency means of each country/economy and other statistics of interest, and to estimate the sampling variability as well as the imputation variance associated with the plausible values.

As described in the earlier section, plausible values are random draws from the posterior distribution of the proficiencies given the item responses  $x_v$ , background variables  $y_v$ , and estimated model parameters. For any student, the value of  $\theta_v$  used in the computation of  $t$  is replaced by a randomly selected value from the student's posterior distribution. Rubin (1987<sup>[43]</sup>) argued 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 (11.10); the variance among them reflects uncertainty due to not

observing  $\theta_v$ . It should be noted that this variance does not include any variability due to sampling from the population.

It cannot be emphasized strongly enough that the plausible values are not a substitute for individual point estimates (e.g., single test scores). Plausible values are used to make accurate group-level inferences, but they should not be used to make any inferences about individuals. Plausible values are only intermediary computations in the calculation of the expectations 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 individual proficiencies with whom they are associated (Marsman et al., 2016<sup>[48]</sup>; von Davier, Gonzalez and Mislevy, 2009<sup>[35]</sup>). Unlike the plausible values, the more familiar ability estimates of educational measurement are optimal for each student (e.g., bias-corrected maximum likelihood estimates, which are consistent estimates of a student's proficiency, or Bayesian posterior mean 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 and biased estimates of population characteristics (Little and Rubin, 1983<sup>[49]</sup>). 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<sup>[41]</sup>).

Once the plausible values for each students have been produced (in *PISA*  $U=10$  plausible values are produced for each student for each domain except Creative Thinking, for which 10 plausible scores are generated<sup>7</sup>), they can be employed to estimate the value of a population, subpopulation or group estimator  $T$  (e.g., mathematics proficiency) and the magnitude of the errors associated with the estimate as follows:

1. Use the vector made up of the of first of the students' plausible values, and calculate the group estimator  $T$  as if the plausible values were the true values of  $\theta$ . Denote the result  $T_1$ .
2. Calculate the sampling variance of  $T_1$ . Denote the result  $V(T_1)$ .

Carry out steps 1 and 2 for each of the  $U$  vectors of plausible values, thus obtaining  $T_u$  and  $V(T_u)$  for  $u = 1, 2, \dots, U$ .

3. The best estimate of the group quantity  $T$  is then the average of  $T_u$ , obtainable from the  $U$  sets of plausible values:

$$T. = \frac{\sum_{u=1}^U T_u}{U}.$$

Formula 11.11

1. An estimate of the error variance of the estimator  $T$  is the sum of two components, which are the variance due to sampling of examinees and the variance due to latency of the proficiency  $\theta$  (often called measurement error):

$$V(T.) = \frac{\sum_{u=1}^U V(T_u)}{U} + \left(1 + \frac{1}{U}\right) \frac{\sum_{u=1}^U (T_u - T.)^2}{U - 1}.$$

Formula 11.12

The first component in  $V(T.)$  reflects uncertainty due to sampling from the population because *PISA* samples only a portion of the entire population of 15-year-old students. The second component reflects uncertainty due to measurement error because the students' proficiencies  $\theta$  are estimated from a limited number of item responses for each respondent.

### **Example for partitioning the estimated error variance**

The following example illustrates the use of plausible values for partitioning the error variance in one country/economy. Annex Table 11.A.17 presents data for six subgroups of students differing in the context questionnaire variable “Books at home” (variable ST013Q01TA, where 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 a domain. This table presents the means  $T_{ug}$  and the sampling standard errors  $V(T_{ug})^{1/2}$  for each plausible value ( $u=1, \dots, 10$ ) and each subgroup defined by the variable ST013Q01TA ( $g=1, \dots, 6$ ). The bottom section of the table shows the resulting estimates and errors for each subgroup.

Because the standard error associated with the group estimator  $T$  is comprised of sampling error and measurement error, it can be reduced by either increasing the precision of the measurement instrument or reducing the sampling error. In PISA, a resampling method is used to estimate the sampling variance  $V(T_{ug})$ , which uses a balanced repeated replication (BRR) approach (See Chapter 10 for details). This component of variance is similar across the ten plausible values; its values are influenced by the homogeneity of proficiencies among students in the subgroup. Note that the sampling error is generally much larger than the measurement error.

### **Application to the PISA 2022 Main Survey**

This section describes the implementation of IRT scaling and population modelling of the PISA 2022 main survey data. Details of the data and procedure implemented, in particular for the mathematics and reading domains that implemented MSAT as well as for the reading fluency items are described first. The dimensionality analyses conducted to verify the applicability of the unidimensional 2PLM and GPCM models to the mathematics MSAT and the innovative creative thinking domains are described next. Then, the country/economy-specific population modelling analyses and the generation of plausible values are detailed. Finally, the procedure utilised to estimate the linking errors between the 2022 and prior PISA cycles is explained.

#### **IRT scaling**

IRT scaling is the first step in the modelling of PISA data. It was conducted through the multi-group IRT calibration and scaling approach described earlier, using the international 2022 main survey data and using the trend item parameters fixed to their values established in the previous PISA cycle (common international or unique country-by-language) to ensure appropriate linking to the PISA scale. Each domain was calibrated separately using the mdltm software (Khorramdel, Shin and von Davier, 2019<sup>[22]</sup>; von Davier, 2005<sup>[23]</sup>) setup to fix already established item parameters and to estimate new ones with the unidimensional 2PLM and GPCM models.

The mathematics and financial literacy assessments included both trend and new items. Reading and science included only trend items. As the innovative domain, creative thinking included only new items. All the PBA and new PBA assessments of mathematics, reading, and science included only trend items, with PBA being the same instruments since 2015 and new PBA being the same instrument as the PISA for Development 2018 instrument (sharing many items in common with PBA) (OECD, 2019<sup>[50]</sup>).

Annex Table 11.A.19 details the number of trend and new items kept in the analyses after some items were dropped due to content and/or psychometric reasons that could not be resolved (1 in mathematics, 1 in reading, 1 in financial literacy and 6 items in creative thinking). Annex Table 11.A.20 the estimation of the new items’ international parameters. However, the unweighted number of item responses was used to check whether the minimum number of 250 responses required for evaluation item-by-country-by-

language interactions (item-fit) was reached. This was done to ensure that the MD and RMSD statistics could be accurately estimated and the decision to estimate unique parameters when item-misfit was detected appropriate. Nonresponses prior to a student's last valid item response in a cluster were considered omitted and treated as incorrect responses; whereas nonresponses at the end of the cluster were considered not-reached and treated as missing. For CBA mathematics and reading, because of their MSAT design, the treatment of omit and not-reached responses was done considering the whole test rather than by cluster.

### ***Estimation of common international and group-specific item parameters***

Different language versions of the assessment used in countries/economies could result in some items functioning differently in some country-by-language groups. Thus, different language versions of the assessment within a country/economy were treated as separate groups when estimating item parameters. In total, 116 country-by-language groups were used in PISA 2022 multiple-group IRT calibrations for CBA reading, mathematics, and science. In creative thinking and financial literacy 102 and 31 country-by-language groups were analysed, respectively. For PBA and New PBA, 4 countries, each using 1 language were analysed.

To account for cultural and language differences, the stepwise calibration process described earlier was implemented to scale the 2022 data. In the first calibration and fit analyses run, for the trend items, common and group-specific item parameter estimates obtained from the PISA 2018 scaling were used as fixed values. For the new items, common item parameters to all the groups were estimated. Given these parameter estimates, RMSD and MD fit statistics were then computed for all items in all groups, and cases with RMSD above a threshold<sup>8</sup> were identified.

In the relatively rare instances where large RMSD misfit was found (values above 0.4), the item was dropped in the specific group (i.e., excluded from scaling in that group). In the subsequent calibrations and fit analyses runs, unique parameters were estimated, as long as there were 250 unweighted responses, gradually lowering the RMSD threshold to 0.12—a value that was found to be optimal for maximizing both the overall model-data fit and the proportion of international item parameters across country-by-language groups (Joo et al., 2019<sup>[51]</sup>). A review of the results obtained in the final calibration run was also conducted to identify any case where even with unique parameters estimated a value below RMSD of 0.18 could not be reached or very low slope parameter (below 0.1) or extreme difficulty parameters (above 5 in absolute value) were obtained. When such cases were found, the item was dropped in the specific group or specific groups.

In addition to ensuring appropriate model fit and reducing the measurement error, maintaining the comparability of scales through common item parameters across countries/economies, assessment modes, and assessments over time is of prime importance. Therefore, the mdlm software used for item calibration implements an algorithm that monitors RMSD and MD across the specified groups and suggests a list of items to be re-estimated for each group. This algorithm seeks to minimize the number of group-specific item parameters needed to fit the data. It does so, item by item, constraining the item parameters to be the same across the groups in which the item exhibits misfit in the same direction (positive or negative). Thus, the same specific item parameters may be unique to one group or multiple groups (e.g., country-by-language groups) exhibiting similar misfit patterns. Ultimately, through the iterative process it may be discovered that the unique parameters common to more than one group need to be relaxed further and re-estimated separately to reach the desired fit. But this is done only when needed so that the total number of unique parameters is minimized across all countries/economies.

## Dimensionality analyses

The results of the scaling analyses just described show that the IRT models used, with the unidimensionality and local independence assumptions, do fit the data quite well. However, it was important to further evaluate these assumptions for the major and the innovative domains, which included a large proportion of newly developed items and all newly developed items.

Residual analyses of field trial mathematics and creative thinking data and residual analysis of main survey creative thinking data were conducted for each country/economy to assess both the conditional independence and unidimensionality assumptions. For mathematics, additional dimensionality analyses of the main survey data were conducted to verify that the new items developed based on the revised framework do not introduce a new dimension, distinct from the one captured by the mathematics PISA scale developed in prior PISA cycles. This was done by fitting a two-dimensional IRT simple-structure model which treated trend and new items as two different latent traits and evaluating the extent to which the more complex two-dimensional model of the total weighted data from all countries/economies provided a significant improvement in fit. These analyses were conducted in the same way as in previous cycles for the major domain of mathematics and the innovative creative thinking domain (OECD, 2017<sup>[3]</sup>; 2020<sup>[52]</sup>). The methods implemented to conduct residual analysis are detailed below; results are reported in the next sections.

The mdltm software (von Davier, 2005<sup>[23]</sup>) computes residuals in the step that follows the item calibration. For dichotomous item responses, response residuals for a person  $v$  with estimated ability  $\hat{\theta}_v$  for each item  $i = 1, \dots, n$  were defined as below:

$$r(x_{vi}) = \frac{x_{vi} - P(X_i = 1 | \hat{\theta}_v)}{\sqrt{P(X_i = 1 | \hat{\theta}_v)[1 - P(X_i = 1 | \hat{\theta}_v)]}} \quad \text{Formula 11.13}$$

For polytomous item responses, response residuals were calculated using the conditional mean and variance defined below:

$$r(x_{vi}) = \frac{x_{vi} - E(X_i | \hat{\theta}_v)}{\sqrt{V(X_i | \hat{\theta}_v)}} \quad \text{Formula 11.14}$$

$$E(X_i | \hat{\theta}_v) = \sum_{k=1}^{m_i} kP(x_{vi} = k | \hat{\theta}_v), \quad \text{Formula 11.15}$$

$$V(X_i | \hat{\theta}_v) = \sum_{k=1}^{m_i} k^2P(x_{vi} = k | \hat{\theta}_v) - [E(X_i | \hat{\theta}_v)]^2. \quad \text{Formula 11.16}$$

Once the item response residuals have been calculated, the item residual correlations across respondents can be computed to produce an item residual correlation matrix. Although the null distribution of such residual correlations--also known as the  $Q_3$  statistic (Yen, 1984<sup>[53]</sup>)—are not well known, unidimensional and locally independent data are expected to show random residual correlations patterns around zero across all items and across items within each unit (Chen and Thissen, 1997<sup>[54]</sup>; Yen, 1984<sup>[53]</sup>). Local item dependencies are found when an item pair shows highly correlated response residuals and their item slope parameter estimates are high. In such cases where an item pair or multiple item pairs within a unit show local item dependence, this may be addressed by scoring these two items or the whole unit as a single

polytomous score and modelled with the partial or generalized partial credit model described earlier in this chapter (Rosenbaum, 1988<sup>[55]</sup>; Wilson and Adams, 1995<sup>[56]</sup>).

Following the inspection of the residual correlation matrix and the treatment of local item dependences, principal component analysis of the residual correlation matrix was conducted to evaluate the extent to which the instrument is unidimensional. If the unidimensionality assumption holds, little common variance among the item response residuals is expected after the ability dimension has been accounted for by the IRT model. In this case, a principal component analysis will produce a scree plot where no single component accounts for much more variance than any other.

### ***Mathematics dimensionality analyses***

Residual-based dimensionality analyses of the CBA mathematics were conducted on the field trial data to identify potential local item dependence and to confirm the unidimensionality of the mathematics instrument assembled for the main survey. Based on the item-by-item correlations for all mathematics items, no item pairs were identified with exceptionally strong correlations. Furthermore, the unidimensional IRT scaling analyses of the field trial data and later the main survey data (as described above) did not show any items with unusually large slope parameters. Both IRT scaling and residual analysis provided evidence that the conditional independence assumption was not violated.

The two-dimensional IRT modelling of the mathematics main survey data, where trend and new items were assigned to two different latent proficiency scales, provided an additional check of the unidimensionality assumption. When the multidimensional IRT model was fitted, the trend item parameters were fixed to the common international item parameters obtained from the PISA 2018 cycle, and the new items were constrained to the newly estimated unidimensional international parameters. Although the Akaike Information Criterion (AIC) (Akaike, 1974<sup>[57]</sup>) showed better fit for the two-dimensional model, the Bayesian Information Criterion (BIC) (Schwarz, 1978<sup>[58]</sup>) and the log-penalty improvement showed that the unidimensional model fits better and the multidimensional model provides very little improvement over the unidimensional model (Annex Table 11.A.21). In particular, it was found that the unidimensional model reached 99.8% of the model fit improvement over the independence model compared to the gains expected from the multidimensional model. Similarly, the two-dimensional IRT model of the field trial data showed only marginal improvement in overall model fit over the unidimensional IRT model. Moreover, the correlations of two sets of group means (the trend item only and the new items only) from the multidimensional model were very high, ranging from 0.91 to 0.99 across the different country-by-language groups. Additionally, the dimension-specific weighted likelihood estimates (WLEs) of student ability were very highly correlated with the unidimensional WLEs.

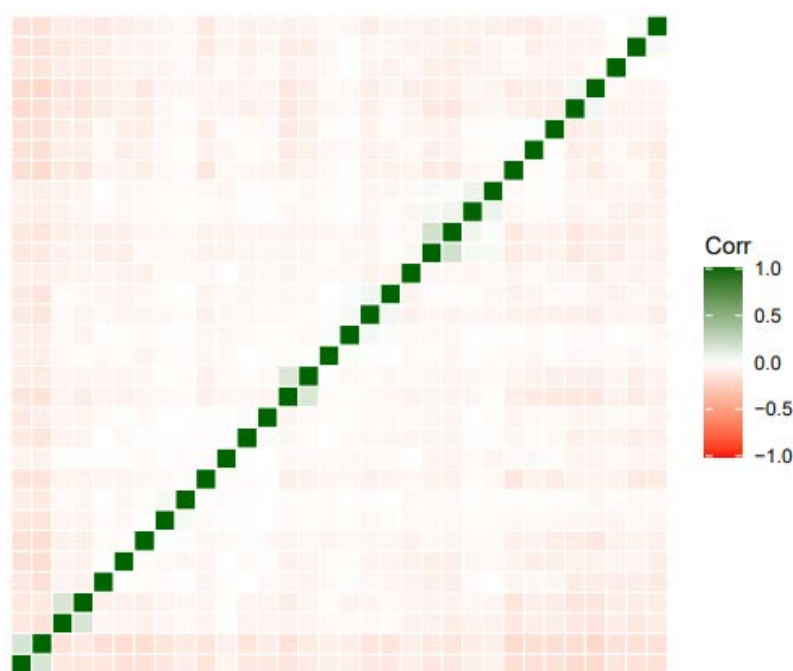
Considering all the evidence gathered from the field trial and main survey data analyses, there is strong evidence that the new and trend mathematics items and scores can be placed on the existing unidimensional PISA scale.

### ***Creative thinking dimensionality analyses***

As the innovative domain, creative thinking was an entirely new domain in 2022. Field trial analyses showed that the instrument was essentially unidimensional. For the main survey, 36 items were selected out of the 40-item field trial item pool. The unidimensional IRT scaling of the main survey data was conducted and response residuals were calculated. Pairwise residual item correlations were then computed for each country-by-language group and averaged across groups. Figure 11.7 shows the residual correlation matrix obtained. Besides the dark green squares on the diagonal that represent each item correlating with itself, no strong pairwise residual correlation and no noticeable patterns that could be indicative of additional dimension(s) was observed.



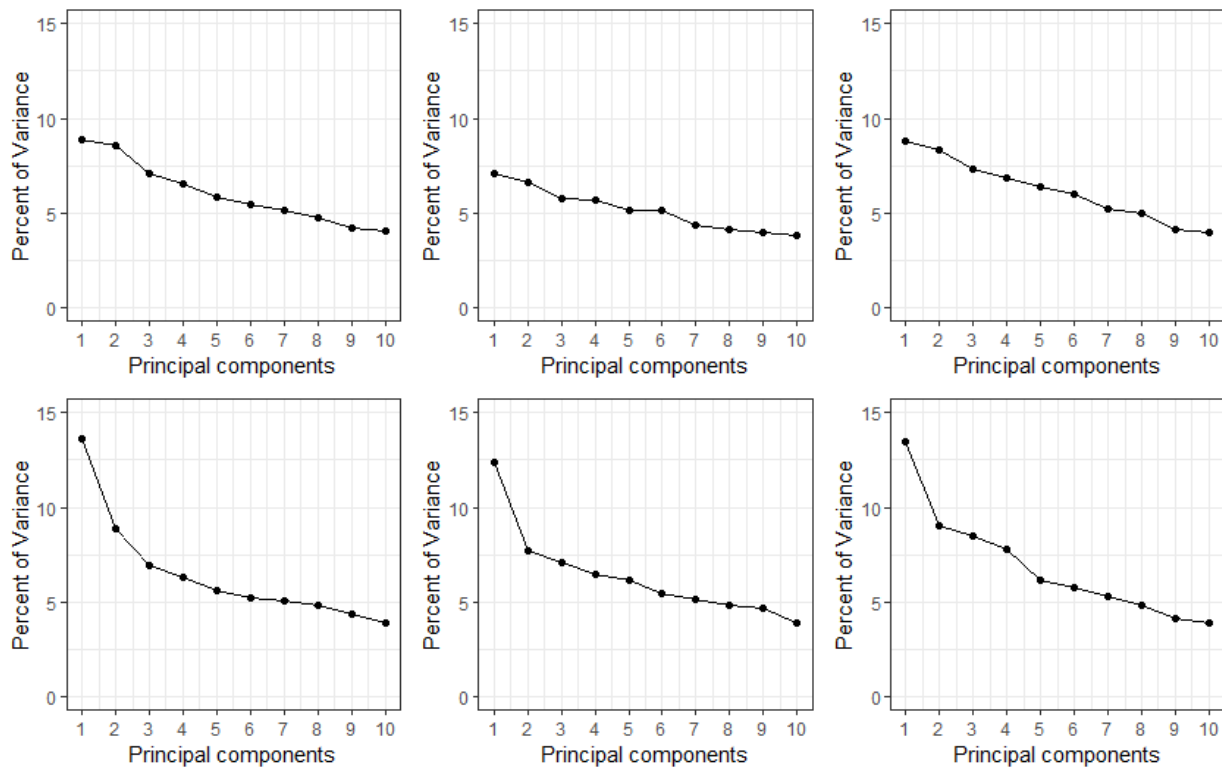
Figure 11.7. Residual correlation matrix for the creative thinking main survey



As part of the residual analysis, principal components of the residual correlation matrix were extracted. Should the eigenvalue of the first principal component be much larger than the other principal components, an additional latent trait, other than the overall ability, could be present. When all the item residual correlations are included as variables, the percentage of variance adds up to 100%. Analysis results across countries/economies, showed that the percentage of variance for the first principal component ranges from 7.1% to 13.7% with a median of 10.2% and the percentage of variance accounted for by the first 10 principal components ranges from 50.8% to 73.65%, with a median value of 63.14%. Thus, the first component did not account for a large part of the variance accounted for by the first ten components. This was confirmed by inspection of each country/economy principal component analysis scree plots in most cases.

The plots in Figure 11.8 show six countries' scree plots as the most distinctive examples. In most cases illustrated by the top three scree plots no clear "elbow" that would be indicative of an additional dimension not accounted for by the unidimensional IRT scaling. However, in a few cases some evidence of multidimensionality was observed. Nevertheless, overall, the results supported the scaling and reporting of creative thinking as proficiency using a unidimensional scale.

Figure 11.8. Percentage of variance from principal component analyses for 6 countries/economies



### **Population modelling in PISA 2022**

The population model described earlier was applied to the PISA 2022 data. Fixing the item parameters to their values obtained from the unidimensional IRT scaling, multivariate latent regression models were fitted to the data at the country/economy level, and 10 plausible values per domain were generated for each student. Plausible values for core domains (reading, mathematics, and science) were generated for all students participating in the assessment, regardless of whether they were administered items in that domain. Plausible values for the innovative domain were generated for all students if countries/economies opted for the CrT domain. That is, students received plausible values for each test domain administered in their country/economy according to the test design implemented regardless of the specific forms they took. Students who did not participate or did not have responses in a particular domain were assigned model-dependent plausible values for that domain based on their responses to the BQ as well as the cognitive responses in other domains.

Measurement errors must be considered when dealing with the plausible values in the secondary analyses. The plausible values for the domain(s) students did not take have larger uncertainty than the plausible values for the other domains that were administered to them. By using repeated analysis with each of the 10 plausible values, the measurement error will readily 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.

While most covariates used in the population modelling come from the student BQ responses, some additional covariates were derived from the cognitive assessment's process data. Same as done in PISA 2018 (see Annex H of the PISA 2018 technical report), such derived covariates include response time information, and school-level WLEs to capture the unique variations across schools, which are relevant for predicting proficiency distributions within each country/economy.

The following sections provide further information about how the population model was applied to PISA 2022 data, how plausible values were generated, and how plausible values can be used in further analyses.

### ***Main sample, creative thinking and financial literacy sample models***

The software called DGROUP (Educational Testing Service, 2012<sup>[59]</sup>) was used to estimate the multivariate latent regression models and generate plausible values (von Davier and Sinharay, 2014<sup>[12]</sup>; von Davier et al., 2006<sup>[36]</sup>). During the estimation, the item parameters for the cognitive items were fixed at the values obtained from the multi-group IRT models described earlier in this chapter. As in previous PISA cycles, nearly all student BQ variables, as well as some contextual characteristics, were included.

All BQ variables were contrast-coded before they were processed further. The contrast coding scheme is reproduced in Annex B of this report. Contrast coding allows for the inclusion of missing responses and avoids the necessity of assuming a linear relationship between the responses to any question and the outcome variable. Note that with the introduction of within-construct matrix sampling design, missing by matrix-sampling design and missing by omitting behavior were distinguished, which increased the number of contrast codes for BQ variables. With contrast-coded BQ variables, a PCA is conducted to 1) remove collinearity among variables when present and 2) reduce the large number of contrast-coded BQ variables into a smaller number of principal components that are sufficient to account for a large proportion of the variation in the BQ variables without over-parameterisation. Because each country/economy can have unique associations among the BQ variables, a set of principal components was calculated for each country/economy. As such, the extraction of principal components was carried out separately by country/economy. In PISA, the number of principal components retained in each of the multivariate latent regression models was selected to be the smaller of 1) the number of principal components needed to explain 80% of the BQ variance, and 2) the number that corresponds to 6.7% (1/15) of the raw sample size. Note that in previous PISA 2015 and 2018 cycles, the number that corresponds to 5% (1/20) of the raw sample size was used. However, with the increase in BQ scales and variables, the rule was relaxed to retain more information in the extracted principal components. Still, this avoided a numerical instability in the estimation that could occur due to potential overparameterization of the model.

The main sample data collection included the core domains administered by all 81 participating countries/economies and the innovative creative thinking domain administered by 64 countries/economies. Separate population modelling analyses of the core domains, of mathematics subscales with reading and science, and of creative thinking with the core domains were conducted. The financial literacy sample data collection was offered as an international option and was administered by 20 countries/economies. The cognitive instruments included trend items from 2012, 2015, and 2018, and a few new items. For the population modelling, the financial literacy sample (who took Forms 67 – 74) was combined with the students from the main sample who took reading and mathematics only (Forms 1 – 12). This was done to establish a stable linkage between the financial literacy and main PISA forms, and the reading and mathematics domains. Thus, the financial literacy sample received plausible values in mathematics, reading, and financial literacy, but not in science and not in mathematics subscales.

### ***Treatment of students with fewer than six test item responses***

This section addresses the issue of students who provided background information but did not respond to enough cognitive items. Students with responses to fewer than six cognitive items in any domain were not included in the multivariate latent regression modelling to avoid unstable estimations of the  $\Gamma$  and  $\Sigma$ .

In PISA 2022, fewer than: 0.09% of students were excluded from the core domains CBA or new PBA multivariate latent regressions; 7.4% the mathematics sub-scales; 0.04% the creative thinking; and less than 0.03% from the financial literacy multivariate latent regressions. Nevertheless, the population model

was applied to these students for the generation of plausible values. For each of the two mathematics subscales (by *process* and by *content*), the proportion of students excluded from the modelling is larger because responses to at least six items in the relatively short subscales were needed to be included in the multivariate latent regression model.

Consistent with the data treatment applied in the IRT scaling, nonresponses prior to a valid response were considered omitted and treated as incorrect responses; whereas nonresponses at the end of each of the cluster (for non-adaptive domains) or each MSAT session (for mathematics and reading) were considered not-reached and treated as missing in the population modelling and PV generation.

### ***Plausible values***

Plausible values for the domains evaluated were drawn from the normal approximations to the posterior distributions estimated from the multivariate latent regression models.

The plausible value variables for the domains follow the naming convention PV1<domain> through PV10<domain>, where “<domain>” took on the following form:

- MATH for mathematics
- READ for reading
- SCIE for science
- CRTH\_NC<sup>9</sup> for creative thinking
- FLIT for financial literacy

### *Population modelling for the mathematics subscales*

The aim of generating plausible values for the different mathematics subscales is to provide proficiency estimates representative of important aspects within the overall mathematics framework. These subscales allow for secondary analyses of relationships between proficiency and BQ variables that focus on different aspects within the mathematics domain. However, it should be noted that subscale proficiencies (plausible values) are based on fewer items than the full scale and, thus, are associated with larger measurement error.

There were two sets of subscales reported for mathematics. These were process subscales related to mathematical reasoning (employing mathematical concepts, facts, and procedures; interpreting, applying, and evaluating mathematical outcomes; formulating situations mathematically; reasoning) and content subscales related to mathematical content knowledge (space and shape; quantity; change and relationships; uncertainty and data). Mathematics subscales were computed for the CBA only. Annex Table 11.A.22 gives an overview of the 233 (one item was dropped) mathematics items by the cognitive process and the test structure. It should be noted that the two mathematics subscale category types are based on a two-way classification of the same 233 items (distributed into the 4 + 4 = 8 subscales). In other words, each item contributed to one of the cognitive process subscales and one of the content subscales.

Because the cognitive process subscales and the content subscales were based on the same set of mathematics items, population modelling for the cognitive process subscales and the population modelling for the content subscales could only be done separately. Therefore, two additional multidimensional population models were fitted for each CBA country/economy to provide the desired mathematics subscale PVs. These two models were:

- Model 1: reading, science, and the four subscales of mathematics cognitive process, thus, 6 dimensions in total;

- Model 2: reading, science, and the four subscales of mathematics content subscales, thus, 6 dimensions in total.

Reading and science data were used for the population modelling of the mathematics subscales to maximize the information used from the students. PVs were generated for those domains (reading and science) in these runs, but only the PVs for the mathematics subscales were included in the database for each set of mathematics subscales.

The item parameters used for the population modelling of the mathematics subscales were the same as those for the overall mathematics scale described above, which were obtained from the unidimensional multi-group IRT model for mathematics. Therefore, the mathematics subscales and the overall mathematics scale proficiencies can be compared as they are on the same scale. However, because the mathematics scale is not the weighted average of the mathematics subscales, a country/economy's mean proficiency in mathematics can be noticeably different from the country/economy's mean subscale proficiencies.

The plausible values reported for the mathematics subscales follow the naming convention PV1<subscale> through PV10<subscale>, where "<subscale>" takes on the following form:

- MCCR Content Subscale of Mathematics – Change and Relationships
- MCQN Content Subscale of Mathematics – Quantity
- MCSS Content Subscale of Mathematics – Space and Shape
- MCUD Content Subscale of Mathematics – Uncertainty and Data
- MPEM Cognitive Process Subscale of Mathematics – Employing Mathematical Concepts, Facts, and Procedures
- MPFS Cognitive Process Subscale of Mathematics – Formulating Situations Mathematically
- MPIN Cognitive Process Subscale of Mathematics – Interpreting, Applying, and Evaluating Mathematical Outcomes
- MPRE Cognitive Process Subscale of Mathematics – Reasoning

Finally, as noted earlier, PVs from the same draw should be used when assessing correlations between domains or when conducting secondary analyses, not from different draws. Thus, estimating correlations between MPEM1, MPFS1, MPIN1, MPRE1 is appropriate, while estimating correlations between MPEM1, MPFS2, MPIN3, MPRE4 is inappropriate. The same is true for the content subscale. Because the core domain PVs and the subscale PVs reported were draws from different population models, estimating correlations between them would not be appropriate. However, the correlations between the other cognitive domains and the subscales that are part of the each one of the two subscale population models estimated are reported in Chapter 14.

### ***Linking PISA 2022 to previous PISA cycles***

There are three measurable sources of error variance to account for when using the PISA data. These are error due to student sampling, error due to the reliability of the assessment, and error due to the linking of different instruments across assessment cycles.

Following the approach implemented in 2015, an evaluation of the magnitude of linking error was conducted by considering differences between reported country/economy results from previous PISA cycles and the transformed results from rescaling prior to 2015. The magnitude of the linking errors is related to the changing assessment framework, instruments, mode of delivery and scaling methods over PISA cycles. It is also related to changes from major to minor domain that could lead to a recombination of items and units within clusters, as well as to changes in design from linear to adaptive.

As in past cycles, scale-level differences across countries/economies between adjacent calibrations are considered as the target of inference. The effect of the variability of two calibrations is evaluated at the cross-country/economy level, while within-country/economy sampling variability is not targeted. Moreover, sampling variance and measurement variance are two separate variance components that are accounted for by the variance estimation based on replicate weights and plausible values. Taken together, the focus of the linking error lies on the expected variability on the country/economy mean over the different calibrations.

The definition of calibration differences starts from the ability estimates of a respondent  $v$  from country/economy  $g$  in a target cycle under two separate calibrations (e.g., the original calibration of a PISA cycle and its recalibration), C1 and C2. We can write for calibration C1:

$$\tilde{\theta}_{v,C1,g} = \theta_{v,true} + \hat{u}_{C1,g} + \tilde{\epsilon}_v,$$

Formula 11.17

Where  $\hat{u}_{C1,g}$  denotes the estimated country/economy specific error term in C1 and  $\tilde{\epsilon}_v$  is the respondent specific measurement error; and for calibration C2 accordingly:

$$\tilde{\theta}_{v,C2,g} = \theta_{v,true} + \hat{u}_{C2,g} + \tilde{\epsilon}_v.$$

Formula 11.18

Defined in this way, there may be country/economy 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/economy-level estimates. Given the assumption of a country/economy-level variability of estimates due to C1 and C2 calibrations, for the differences between estimates we find:

$$\tilde{\theta}_{v,C1,g} - \tilde{\theta}_{v,C2,g} = \hat{u}_{C1,g} - \hat{u}_{C2,g},$$

Formula 11.19

and the expectation can be estimated by:

$$E(\hat{u}_{C1,g} - \hat{u}_{C2,g}) = \tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2} = \hat{\Delta}_{C1,C2,g}.$$

Formula 11.20

Across countries/economies, the expected differences of country/economy means ( $\tilde{\mu}$ ) can be assumed to vanish, since the scales are transformed after calibrations to match distribution moments. That is, we may assume:

$$\sum_{g=1}^G E(\hat{u}_{C1,g} - \hat{u}_{C2,g}) = 0 = \sum_{g=1}^G \hat{\Delta}_{C1,C2,g}.$$

Formula 11.21

The variance of the differences of country/economy 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 linking error can be written as:

$$V[\hat{\Delta}_{C1,C2,g}] = \frac{1}{G} \sum_{g=1}^G (\tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2})^2.$$

Formula 11.22

The main characteristics of this approach can be summarised as follows:

- Scale-level differences across countries/economies 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/economy level.
- Within-country/economy 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 the formula (11.22) 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. To avoid the possibility that some data points (countries/economies) have excessive influence on the results, the robust  $S_n$  statistic was used, as it was in PISA 2015 and 2018. The  $S_n$  statistic was proposed by Rousseeuw and Croux (1993<sup>[2]</sup>) as a more efficient alternative to the scaled median absolute deviation from the median ( $1.4826 \cdot \text{MAD}$ ) that is commonly used as a robust estimator of standard deviation. It is defined as:

$$S_n = 1.1926 * \text{med}_i \left( \text{med}_j (|x_i - x_j|) \right).$$

Formula 11.23

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 14.

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>.

**Table 11.1. Detailed analysis: Scaling PISA data**

Table	Title
Figure 11.1	Main sample yield for countries/economies participating in the CBA
Figure 11.2	Financial literacy sample yield for participating countries/economies
Figure 11.3	Main sample yield for countries/economies participating in the PBA and new PBA
Web Figure 11.4a	Proportion of students routed to each testlet combination in mathematics MSAT
Web Figure 11.4b	Proportion of students routed to each testlet combination in reading MSAT
Figure 11.5	Mathematics median response time by median proficiency across countries/economies
Figure 11.6	Distribution of mathematics response time in each country/economy
Web Figure 11.7	Median item response time by proficiency level for mathematics trend items
Web Figure 11.8	Median item response time by proficiency level for mathematics new items
Web Figure 11.9	Median item response times for creative thinking items
Figure 11.10	Item response curve (ICC) for an item where the common item parameter is not appropriate for one group
Figure 11.11	Residual correlation matrix for the creative thinking main survey
Figure 11.12	Percentage of variance from principal component analyses for 6 countries/economies

StatLink  <https://stat.link/5au4gy>

## References

- Akaike, H. (1974), “A new look at the statistical model identification”, *IEEE Transactions on Automatic Control*, Vol. 19/6, pp. 716–723, <https://doi.org/10.1109/TAC.1974.1100705>. [57]
- Beaton, A. (ed.) (1987), *Joint estimation procedures*, Educational Testing Service. [37]
- Birnbaum, A. (1968), “Some latent trait models and their use in inferring an examinee’s ability”, in Lord, F. and M. Novick (eds.), *Statistical Theories of Mental Test Scores*, Addison-Wesley. [6]
- Bock, R. and M. Aitkin (1981), “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm”, *Psychometrika*, Vol. 46/4, pp. 443-459. [24]
- Bock, R. and M. Zimowski (1997), “Multiple group IRT”, in van der Linden, W. and R. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer-Verlag. [4]
- Chang, H. and W. Stout (1993), “The asymptotic posterior normality of the latent trait in an IRT model”, *Psychometrika*, Vol. 58/1, pp. 37 - 52, <https://doi.org/10.1007/BF02294469>. [44]
- Chen, W. and D. Thissen (1997), “Local dependence indexes for item pairs using item response theory”, *Journal of Educational and Behavioral Statistics*, Vol. 22/3, pp. 265–289, <https://doi.org/10.3102/10769986022003265>. [54]
- Dempster, A., N. Laird and D. Rubin (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm”, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39/1, pp. 1-38, <https://www.jstor.org/stable/2984875>. [42]
- Educational Testing Service (2012), *DGROUP [Computer software]*. [59]
- Fischer, G. and I. Molenaar (eds.) (1995), *Rasch Models: Foundations, Recent Developments, and Applications*, Springer. [9]
- Glas, C. (2010), “Item Parameter Estimation and Item Fit Analysis”, in van der Linden, W. and C. Glas (eds.), *Elements of Adaptive Testing*, Springer. [27]
- Glas, C. and K. Jehangir (2014), “Modelling Country Specific Differential Item Functioning”, in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of International Large-scale Assessment*, CRC Press. [18]
- Glas, C. and N. Verhelst (1995), “Testing the Rasch Model”, in Fischer, G. and I. Molenaar (eds.), *Rasch Models: Foundations, Recent Developments, and Applications*, Springer. [16]
- Jewsbury, P. et al. (2023), *Modeling multistage and targeted testing data with item response theory*, [Manuscript submitted for publication], Research and Development Division, Educational Testing Service. [30]



- Jewsbury, P. and P. van Rijn (2020), “IRT and MIRT models for item parameter estimation with multidimensional multistage tests”, *Journal of Educational and Behavioral Statistics*, Vol. 45/4, pp. 383-402. [26]
- Johnson, E. (1989), “Considerations and techniques for the analysis of NAEP data”, *Journal of Educational Statistics*, Vol. 14/4, pp. 303-334, <https://doi.org/10.3102/10769986014004303>. [45]
- Johnson, E. and K. Rust (1992), “Population inferences and variance estimation for NAEP data”, *Journal of Educational Statistics*, Vol. 17/2, pp. 175–190, <https://doi.org/10.3102/10769986017002175>. [46]
- Joo, S. et al. (2019), *Evaluating Item Fit Statistic Thresholds in PISA: The Analysis of Cross-Country Comparability of Cognitive Items*, [Manuscript submitted for publication], Research and Development Division, Educational Testing Service. [51]
- Khorramdel, L., H. Shin and M. von Davier (2019), “GDM software mdltm including parallel EM algorithm”, in von Davier, M. and Y. Lee (eds.), *Handbook of Psychometric Models for Cognitive Diagnosis*, Springer. [22]
- Kirsch, I. et al. (2013), “On the growing importance of international large-scale assessments”, in von Davier, M., E. Gonzalez and I. Kirsch (eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, Springer, [https://doi.org/10.1007/978-94-007-4629-9\\_1](https://doi.org/10.1007/978-94-007-4629-9_1). [32]
- 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. 49/4, pp. 764–766, <https://doi.org/10.1016/j.jesp.2013.03.013>. [1]
- Little, R. and D. Rubin (1983), “On jointly estimating parameters and missing data”, *American Statistician*, Vol. 37/3, pp. 218–220. [49]
- Marsman, M. et al. (2016), “What can we learn from plausible values?”, *Psychometrika*, Vol. 81/2, pp. 274–289, <https://doi.org/10.1007/s11336-016-9497-x>. [48]
- Meredith, W. (1993), “Measurement invariance, factor analysis and factorial invariance”, *Psychometrika*, Vol. 68/4, pp. 525–543, <https://doi.org/10.1007/BF02294825>. [14]
- Meredith, W. and J. Teresi (2006), “An essay on measurement and factorial invariance”, *Medical Care*, Vol. 44/11, pp. S69–S77, <https://doi.org/10.1097/01.mlr.0000245438.73837.89>. [19]
- Mislevy, R. (1991), “Randomization-based inference about latent variables from complex samples”, *Psychometrika*, Vol. 56/2, pp. 177–196, <https://doi.org/10.1007/BF02294457>. [33]
- Mislevy, R. et al. (1992), “Estimating population characteristics from sparse matrix samples of item responses.”, *Journal of Educational Measurement*, Vol. 29/2, pp. 133–161, <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>. [41]
- Mislevy, R. and K. Sheehan (1987), “Marginal Estimation Procedures”, in Beaton, A. (ed.), *Implementing the New Design: The NAEP 1983-84 Technical Report*, Educational Testing Service. [39]
- Muraki, E. (1992), “A generalized partial credit model: Application of an EM algorithm”, *Applied Psychological Measurement*, Vol. 16/2, pp. 159-177, <https://doi.org/10.1002/j.2333-8504.1992.tb01436.x>. [7]

- OECD (2022), *PISA 2018 Technical Report*. [31]
- OECD (2020), *PISA 2018 Technical Report*, PISA, OECD Publishing, Paris, [52]  
<https://www.oecd.org/pisa/data/pisa2018technicalreport/>.
- OECD (2019), *PISA for Development Technical Report*, OECD Publishing, Paris, [50]  
<http://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport/>.
- OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, [3]  
<http://www.oecd.org/pisa/data/2015-technical-report/>.
- Oliveri, M. 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, <https://doi.org/10.1080/15305058.2013.825265>. [21]
- Oliveri, M. and M. von Davier (2011), “Investigation of model fit and score scale comparability in international assessments”, *Psychological Test and Assessment Modelling*, Vol. 53/3, pp. 315–333. [20]
- Patz, R. and B. Junker (1999), “A straightforward approach to Markov chain Monte Carlo methods for item response models”, *Journal of Educational and Behavioral Statistics*, Vol. 24/2, pp. 146 - 178, <https://doi.org/10.2307/1165199>. [40]
- Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Nielsen and Lydiche. [8]
- Reise, S., K. Widaman and R. Pugh (1993), “Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance”, *Psychological Bulletin*, Vol. 114/3, pp. 552–566, <https://doi.org/10.1037/0033-2909.114.3.552>. [15]
- Rosenbaum, P. (1988), “Permutation tests for matched pairs with adjustments for covariates.”, *Applied Statistics*, Vol. 37/3, pp. 401–411, <https://doi.org/10.2307/2347314>. [55]
- Rousseeuw, P. and C. Croux (1993), “Alternatives to the median absolute deviation”, *Journal of the American Statistical Association*, Vol. 88/424, pp. 1273–1283, <https://doi.org/10.2307/2291267>. [2]
- Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons. [43]
- Rust, K. (2014), “Sampling, weighting, and variance estimation in international large-scale assessments”, in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press. [47]
- Rutkowski, L., M. von Davier and D. Rutkowski (eds.) (2014), *Analytics in International Large-Scale Assessments: Item Response Theory and Population Models*, CRC Press. [12]
- Schwarz, G. (1978), “Estimating the Dimension of a Model”, *Annals of Statistics*, Vol. 6/2, pp. 461 - 464. [58]
- Thomas, N. (2002), “The role of secondary covariates when estimating latent trait population distributions”, *Psychometrika*, Vol. 67/1, pp. 33–48, <https://doi.org/10.1007/BF02294708>. [34]

- Thomas, N. (1993), "Asymptotic corrections for multivariate posterior moments with factored likelihood functions", *Journal of Computational and Graphical Statistics*, Vol. 2/3, pp. 309–322. [38]
- van der Linden, W. and R. Hambleton (eds.) (2016), *Handbook of Modern Item Response Theory*, Springer. [11]
- van der Linden, W. and R. Hambleton (1997), "Item Response Theory: Brief History, Common Models, and Extensions", in van der Linden, W. and R. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer. [10]
- van Rijn, P. and H. Shin (2019), *Item Calibration for Multistage Tests in the Context of Large-Scale Educational Assessment*, [Manuscript in preparation], Research and Development Division, Educational Testing Service. [29]
- von Davier, M. (2005), "A general diagnostic model applied to language testing data", *Research Report No. RR-05-16*, Educational Testing Service. [23]
- von Davier, M., E. Gonzalez and R. Mislevy (2009), "What are plausible values and why are they useful?", *IERI Monograph Series*, No. 2/1, [http://www.ierinstitute.org/IERI\\_Monograph\\_Volume\\_02\\_Chapter\\_01.pdf](http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf). [35]
- von Davier, M. et al. (2006), "Statistical Procedures Used in the National Assessment of Educational Progress (NAEP): Recent Developments and Future Directions", in Rao, C. and S. Sinharay (eds.), *Handbook of Statistics: Psychometrics*, Elsevier, [https://doi.org/10.1016/S0169-7161\(06\)26032-2](https://doi.org/10.1016/S0169-7161(06)26032-2). [36]
- 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, <https://doi.org/10.1177/0146621604268734>. [5]
- von Davier, M. et al. (2019), "Evaluating item response theory linking and model fit for data from PISA 2000-2012", *Assessment in Education: Principles, Policy & Practice*, Vol. 26/4, pp. 466-488, <https://doi.org/10.1080/0969594X.2019.1586642>. [13]
- Wilson, M. and R. Adams (1995), "Rasch models for item bundles", *Psychometrika*, Vol. 60/2, pp. 181–198, <https://doi.org/10.1007/BF02301412>. [56]
- Xu, X. and M. von Davier (2008), "Comparing multiple-group multinomial log-linear models for multidimensional skill distributions in the general diagnostic model", *ETS Research Report Series*, No. ETS RR-08-35, Educational Testing Service, Princeton, NJ. [25]
- Yamamoto, K. (1997), "Scaling and scale linking", in Murray, T., I. Kirsch and L. Jenkins (eds.), *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey*, National Center for Education Statistics. [17]
- Yamamoto, K. et al. (forthcoming), "Improved test designs and multistage adaptive testing in large-scale assessments", in Khorramdel, L. et al. (eds.), *Innovative Computer-Based International Large-Scale Assessments: Foundations, Methodologies and Quality Assurance Procedures*, Springer. [28]
- Yen, W. (1984), "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model", *Applied Psychological Measurement*, Vol. 8/2, pp. 125–145, <https://doi.org/10.1177/014662168400800201>. [53]

## Notes

---

1. More detail on the parallel MSAT Reading and Mathematics can be found in Chapter 2 of this Technical Report.
2. With MSAT, testlets of different difficulty are assembled specifically for each stage (core, stage 1 and stage 2), therefore position effects cannot easily be compared across stages.
3. Computed using senate weights so that all countries/economies contribute equally.
4. Note that the parameterisations  $(\theta_v - b_i + d_{ir})$  and  $(\theta_v - b_{ir})$ , both used in the IRT literature, are equivalent. However, the former has the advantage of using  $b_i$  with both the 2PLM and GPCM, representing the overall item difficulty.
5. In contrast, tests that are used to report individual-level results are concerned with accurately assessing the performance of each individual test-taker for the purposes of diagnosis, selection, or placement. This is achieved by administering a relatively large number of items to each individual, resulting in a negligible level of uncertainties associated with the point estimates.
6. The contrast variables derived from the BQ responses can be found in the Annex B to this Technical Report
7. As the mathematical properties of both plausible values and scores (the latter being obtained via a non-linear transformation of the former), plausible values will be used throughout the chapter for brevity.
8. Note that RMSD are always larger than absolute MD values. Therefore, unless one wishes to set different thresholds on RMSD and MD to identify misfit, it is sufficient to use a single threshold on RMSD.
9. Population modeling and plausible values are first produced on each domain's IRT theta scale and then transformed to each domain's reported PISA scale. All domains other than creative thinking use a linear transformation. Creative thinking uses a non-linear test characteristic curve transformation that results in plausible values that correspond to the student's plausible number correct (NC) on a form made up of all the items in the creative thinking item pool.

# Annex 11.A. Detailed Procedures and Techniques

**Annex Table 11.A.1. Chapter 11: PISA Assessment Data: Detailed Analyses**

Tables	Title
Table 11.A.2	Language(s) of assessment, mode of assessment, and number of students and schools sampled for each country/economy
Table 11.A.3	Example output for examining response distributions
Table 11.A.4	Example table of item score category analysis and item flags summary
Table 11.A.5	Flagging criteria for items in the item analyses
Table 11.A.6	Percentage of response time outliers by domain
Table 11.A.7	Descriptive statistics for testlet or cluster response time (in minutes)
Table 11.A.8.	Descriptive statistics for domain stage response time (in minutes)
Web Table 11.A.9	Median domain response time (in minutes) by proficiency level
Table 11.A.10	Median response time (in minutes) by cluster position in the CBA for non-adaptive domains
Table 11.A.11	Median response time (in minutes) by assessment hour in the CBA for all domains
Table 11.A.12	Average proportion correct (P+) by cluster position in the CBA for non-adaptive domains
Table 11.A.13	Average proportion correct (P+) by assessment hour in the CBA for all domains
Table 11.A.14	Average proportion of omitted responses by cluster position in the CBA for non-adaptive domains
Table 11.A.15	Average omission rate by assessment hour in the CBA for all domains
Table 11.A.16	Average proportion correct (P+) by cluster position in new PBA
Table 11.A.17	Average proportion of omitted responses by cluster position in new PBA
Table 11.A.18	Example for use of plausible values for partitioning the error
Table 11.A.19	Panel Two of Example for use of plausible values for partitioning the error
Table 11.A.20	Number of trend (linking) items and new items by domain and mode of assessment
Table 11.A.21	Unweighted calibration sample size by domain and mode of assessment
Table 11.A.22	Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new mathematics items in the main survey
Table 11.A.23	Distribution of the items to the mathematics subscales

StatLink  <https://stat.link/4bkjxo>

**Annex Table 11.A.2. Language(s) of assessment, mode of assessment, and number of students and schools sampled for each country/economy**

Country	Language(s)	Test Mode	Main Sample	Financial Literacy Sample	Total	Schools
Albania (ALB)	Albanian	CBA	6,156		6,156	283
Argentina (ARG)	Spanish	CBA	12,127		12,127	460
Australia (AUS)	English	CBA	13,521		13,521	761
Austria (AUT)	German	CBA	6,159	1,599	7,758	304
Baku (Azerbaijan) (QAZ)	Azeri, Russian	CBA	7,720		7,720	199
Belgium* (BEL)	French, German, Dutch	CBA	8,286	1,189	9,475	285
Brazil (BRA)	Portuguese	CBA	10,810	2,901	13,711	602
Brunei Darussalam (BRN)	English	CBA	5,576		5,576	54
Bulgaria (BGR)	Bulgarian	CBA	6,118	1,605	7,723	203
Cambodia (KHM)	Khmer	New PBA	5,279		5,279	183
Canada* (CAN)	French, English	CBA	23,386	4,203	27,589	885
Chile (CHL)	Spanish	CBA	6,489		6,489	231
Chinese Taipei (TAP)	Chinese	CBA	5,896		5,896	188
Colombia (COL)	Spanish	CBA	7,804		7,804	262
Costa Rica (CRI)	Spanish	CBA	6,122	1,453	7,575	199
Croatia (HRV)	Croatian	CBA	6,135		6,135	180
Cyprus (QCY)	Greek, English	CBA	6,517		6,517	102
Czech Republic (CZE)	Czech	CBA	8,460	2,213	10,673	430
Denmark (DNK)	Danish, Faroese	CBA	6,224	1,578	7,802	349
Dominican Republic (DOM)	Spanish	CBA	6,902		6,902	254
El Salvador (SLV)	Spanish	CBA	6,705		6,705	290
Estonia (EST)	Russian, Estonian	CBA	6,392		6,392	196
Finland (FIN)	Finnish, Swedish	CBA	10,256		10,256	242
France (FRA)	French	CBA	6,771		6,771	283
Georgia (GEO)	Georgian, Azerbaijani, Russian	CBA	6,583		6,583	267
Germany (DEU)	German	CBA	7,712		7,712	259
Greece (GRC)	Greek	CBA	6,545		6,545	235
Guatemala (GTM)	Spanish	New PBA	5,190		5,190	290
Hong Kong (China) (HKG)	Chinese, English	CBA	6,048		6,048	168
Hungary (HUN)	Hungarian	CBA	6,236	1,639	7,875	263
Iceland (ISL)	Icelandic	CBA	3,367		3,367	136
Indonesia (IDN)	Indonesian	CBA	13,471		13,471	412
Ireland (IRL)	Irish, English	CBA	5,569		5,569	170
Israel (ISR)	Hebrew, Arabic	CBA	6,251		6,251	193
Italy (ITA)	Italian, German	CBA	10,564	2,789	13,353	345
Jamaica (JAM)	English	CBA	3,956		3,956	154
Japan (JPN)	Japanese	CBA	5,760		5,760	182
Jordan (JOR)	Arabic	CBA	7,799		7,799	260
Kazakhstan (KAZ)	Kazakh, Russian	CBA	19,768		19,768	571
Korea (KOR)	Korean	CBA	6,454		6,454	186
Kosovo (KSV)	Serbian, Albanian	CBA	6,027		6,027	229
Latvia (LVA)	Latvian, Russian	CBA	5,394		5,394	226
Lithuania (LTU)	Lithuanian, Russian, Polish	CBA	7,257		7,257	292
Macao (China) (MAC)	English, Chinese, Portuguese	CBA	4,384		4,384	46
Malaysia (MYS)	Malay, English	CBA	7,069	1,818	8,887	199

Country	Language(s)	Test Mode	Main Sample	Financial Literacy Sample	Total	Schools
Malta (MLT)	Maltese, English	CBA	3,127		3,127	46
Mexico (MEX)	Spanish	CBA	6,288		6,288	280
Mongolia (MNG)	Kazakh, Mongolian	CBA	6,999		6,999	195
Montenegro (MNE)	Montenegrin, Albanian	CBA	5,800		5,800	64
Morocco (MAR)	French, Arabic	CBA	6,867		6,867	178
Netherlands (NLD)	Dutch	CBA	5,046	1,278	6,324	154
New Zealand (NZL)	English	CBA	4,830		4,830	175
North Macedonia (MKD)	Macedonian, Albanian	CBA	6,610		6,610	111
Norway (NOR)	Nynorsk, Bokmål	CBA	6,616	1,719	8,335	266
Palestinian Authority (PSE)	Arabic, English	CBA	7,905		7,905	273
Panama (PAN)	Spanish, English	CBA	4,590		4,590	227
Paraguay (PRY)	Spanish	New PBA	5,087		5,087	283
Peru (PER)	Spanish	CBA	6,968	1,819	8,787	336
Philippines (PHL)	English	CBA	7,193		7,193	188
Poland (POL)	Polish	CBA	6,048	1,574	7,622	246
Portugal (PRT)	Portuguese	CBA	6,819	1,805	8,624	226
Qatar (QAT)	Arabic, English	CBA	7,676		7,676	229
Republic of Moldova (MDA)	Russian, Romanian	CBA	6,235		6,235	265
Romania (ROU)	Romanian, Hungarian	CBA	7,364		7,364	262
Saudi Arabia (SAU)	Arabic, English	CBA	6,928	1,829	8,757	193
Serbia (SRB)	Hungarian, Serbian	CBA	6,432		6,432	185
Singapore (SGP)	English	CBA	6,608		6,608	165
Slovak Republic (SVK)	Slovak, Hungarian	CBA	5,833		5,833	289
Slovenia (SVN)	Slovenian	CBA	6,752		6,752	350
Spain (ESP)	Catalan, Galician, Basque, Spanish, Valencian	CBA	30,920	1,682	32,602	983
Sweden (SWE)	Swedish, English	CBA	6,079		6,079	263
Switzerland (CHE)	German, French, Italian	CBA	6,847		6,847	262
Thailand (THA)	Thai	CBA	8,507		8,507	280
Türkiye (TUR)	Turkish	CBA	7,250		7,250	196
Ukrainian regions (QUR)	Ukrainian	CBA	4,005		4,005	176
United Arab Emirates (ARE)	Arabic, English	CBA	24,623	6,452	31,075	843
United Kingdom (Excl. Scotland) (QUK)	Welsh, English	CBA	9,932		9,932	345
United Kingdom (Scotland) (QSC)	English	CBA	3,277		3,277	120
United States (USA)	English	CBA	4,602	1,121	5,723	160
Uruguay (URY)	Spanish	CBA	6,747		6,747	230
Uzbekistan (UZB)	Karakalpak, Uzbek, Russian	CBA	7,293		7,293	202
Viet Nam (VNM)	Vietnamese	PBA	6,137		6,137	180

Note: Ukrainian regions (QUR) - 18 out of 27 regions administered the assessment.

\*Denotes a country/economy for which the financial literacy domain was not fully sampled across the population; it is not a nationally-representative sample.

## Annex Table 11.A.3. Example output for examining response distributions

BLOCK M01 (UNWEIGHTED)  
Response Analysis

Which plan best represents the drawing o									
	1	NOT RCH	OFF TSK	OMIT	0	1		TOTAL	R BIS =
ITEM 1	N	1	14	74	2054	5466		7608	PT BIS = 0.4551
	PERCENT	0.01	0.18	0.97	27.00	71.85		100.00	P+ = 0.7185
CM033Q01S	MEAN SCORE	7.00	5.00	1.22	3.59	7.31		6.25	DELTA = 10.69
	STD. DEV.	0.00	3.09	1.87	2.92	3.49		3.75	
MAC	RESP WT	0.00	0.00	0.00	0.00	1.00			ITEM WT = 1.00

Which is the third fastest time?									
	2	NOT RCH	OFF TSK	OMIT	0	1		TOTAL	R BIS =
ITEM 2	N	8	0	98	2204	5299		7601	PT BIS = 0.4722
	PERCENT	0.11	0.00	1.29	29.00	69.71		100.00	P+ = 0.6971
CM474Q01S	MEAN SCORE	1.25	0.00	1.38	3.66	7.42		6.25	DELTA = 10.94
	STD. DEV.	1.48	0.00	1.66	3.06	3.14		3.75	
MAC	RESP WT	0.00	0.00	0.00	0.00	1.00			ITEM WT = 1.00

How many people (boys and girls combined)												
	3	NOT RCH	OFF TSK	OMIT	00	11	12	13	21		TOTAL	R BIS =
ITEM 3	N	20	1	1139	1639	335	530	201	3744		7589	PT BIS = 0.7118
	PERCENT	0.26	0.01	15.04	15.01	4.41	6.98	2.65	49.33		100.00	P+ = 0.5636
DM155Q02C	MEAN SCORE	1.00	3.00	2.58	2.58	5.40	5.81	6.33	8.81		6.26	DELTA = 12.36
	STD. DEV.	0.55	0.00	2.02	2.02	2.48	2.69	2.71	2.84		3.74	
HUM	RESP WT	0.00	0.00	0.00	0.00	0.50	0.50	0.50	1.00			ITEM WT = 2.00

## Annex Table 11.A.4. Example table of item score category analysis and item flags summary

BLOCK M01 (UNWEIGHTED)  
Item Score Category Analysis (Partial credit model)

	Category	N	Pct. At	Pct. Below	Mean	Std. Dev.	Biserial	B *
ITEM 1	0	2142	28.15	0.00	3.52	2.93		
CM033Q01S	1	5466	71.85	28.15	7.31	3.49	0.6064	-0.9529
ITEM 2	0	2302	30.29	0.00	3.57	3.05		
CM474Q01S	1	5299	69.71	30.29	7.42	3.14	0.6213	-0.8303
ITEM 3	0	2779	36.62	0.00	3.01	2.31		
DM155Q02C	1	1066	14.05	36.62	5.78	2.65	0.6114	0.3033
	2	3744	49.33	50.67	8.81	2.84	0.5728	-0.8367

BLOCK M01 (UNWEIGHTED)  
Item Analysis Flag Summary

Item ID	Num Resp	Type	R-BIS	P-PLUS	% NOTRCH	% OFFTSK	% OMIT	% MISS	Flags
CM033Q01	2	SCR	0.6064	0.7185	0.01	0.18	0.97	1.17	.....
CM474Q01	2	SCR	0.6213	0.6971	0.11	0.00	1.29	1.39	.....
DM155Q02	5	ECR	0.8431	0.5636	0.26	0.01	15.01	15.25	...O..

## Annex Table 11.A.5. Flagging criteria for items in the item analyses

	Criteria for flagging items
min rbis/rpoly	0.3
min P+	0.2
max P+	0.9
max Omit%	10
max Offtask%	10
max Not-Reached%	10



**Annex Table 11.A.6. Percentage of response time outliers by domain**

DOMAIN	Reading	Science	Mathematics	Financial Literacy	Creative Thinking
Number of Clusters/testlets	30 MSAT testlets	6	144 MSAT testlets	2	5
Number of Outliers	0.68%	1.21%	0.95%	0.53%	0.76%

Note: Statistics for mathematics, reading, science, and creative thinking are based on the main sample; statistics for financial literacy are based on the financial literacy sample.

**Annex Table 11.A.7. Descriptive statistics for testlet or cluster response time (in minutes)**

DOMAIN	N	MIN	Q1	MEDIAN	Q3	MAX	MEAN	SD
Math Testlet 1	543,174	0.04	11.86	16.32	21.31	33.94	16.64	6.80
Math Testlet 2	556,894	0.02	9.86	14.02	17.97	33.94	13.88	5.86
Math Testlet 3	541,703	0.01	6.48	10.17	13.63	33.90	10.16	5.05
Reading Testlet 1	238,303	0.04	9.90	13.64	17.91	31.85	14.04	6.24
Reading Testlet 2	241,238	0.05	13.11	17.67	22.04	37.78	17.44	6.73
Reading Testlet 3	232,806	0.00	6.35	10.44	14.04	37.61	10.29	5.29
Science	236,767	0.03	15.49	21.35	27.47	48.49	21.82	9.50
Financial Literacy	41,682	0.03	15.87	21.90	30.40	53.79	23.22	10.88
Creative Thinking	143,429	0.07	13.60	18.84	24.47	43.21	19.25	8.25

Note: Statistics for mathematics, reading, science, and creative thinking are based on the main sample; statistics for financial literacy are based on the financial literacy sample.

**Annex Table 11.A.8. Descriptive statistics for domain stage response time (in minutes)**

DOMAIN	N	MIN	Q1	MEDIAN	Q3	MAX	MEAN	SD
Mathematics Linear	136,377	0.04	32.23	43.07	50.43	91.06	40.27	12.99
Mathematics MSAT	406,552	0.04	32.52	43.49	50.71	93.85	40.58	12.97
Reading Design A	178,444	0.04	34.91	44.72	50.41	94.60	41.45	12.48
Reading Design B	59,183	0.09	35.02	44.86	50.34	94.06	41.41	12.57
Reading Design A (with RF)	173,578	0.04	34.92	44.76	50.45	94.60	41.46	12.48
Reading Design B (with RF)	57,601	0.09	35.07	44.92	50.38	94.06	41.45	12.57
Science	236,767	0.05	36.17	46.65	53.70	101.34	43.73	12.88
Financial Literacy	41,682	0.11	40.92	49.10	53.32	104.69	45.84	11.57
Creative Thinking	143,429	0.06	30.08	40.35	48.47	93.46	38.52	12.61

Note: Statistics for mathematics, reading, science, and creative thinking are based on the main sample; statistics for financial literacy are based on the financial literacy sample.

**Annex Table 11.A.9. Median response time (in minutes) by cluster position in the CBA for non-adaptive domains**

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Science	29.69	18.08	23.87	17.77	-11.91
Financial Literacy	32.97	16.82	27.38	17.55	-15.42
Creative Thinking	23.81	17.38	20.43	15.98	-7.83

Note: Excludes cluster outliers.

**Annex Table 11.A.10. Median response time (in minutes) by assessment hour in the CBA for all domains**

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Math Linear	45.65	40.03	-5.61
Math MSAT	46.05	40.43	-5.62
Reading Core Items	15.17	12.29	-2.89
Reading Stage 1 and 2	29.44	28.08	-1.37
Reading MSAT	46.75	42.12	-4.64
Science	49.9	42.98	-6.92
Financial Literacy	50.46	46.92	-3.54
Creative Thinking	42.96	37.61	-5.35

Note: Excludes cluster outliers.

**Annex Table 11.A.11. Average proportion correct (P+) by cluster position in the CBA for non-adaptive domains**

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Science	0.452	0.399	0.423	0.384	-0.068
Financial Literacy	0.510	0.434	0.479	0.422	-0.089
Creative Thinking	0.479	0.453	0.456	0.428	-0.051

**Annex Table 11.A.12. Average proportion correct (P+) by assessment hour in the CBA for all domains**

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Math Linear - trend*	0.376	0.357	-0.019
Math Linear - new*	0.388	0.375	-0.013
Math MSAT - trend*	0.376	0.356	-0.02
Math MSAT - new*	0.397	0.385	-0.012
Reading Core Items	0.569	0.524	-0.044
Reading Stage 1 and 2	0.495	0.474	-0.021
Creative Thinking	0.466	0.442	-0.024

**Annex Table 11.A.13. Average proportion of omitted responses by cluster position in the CBA for non-adaptive domains**

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Science	0.027	0.049	0.042	0.06	0.033
Financial Literacy	0.027	0.063	0.041	0.073	0.045
Creative Thinking	0.042	0.041	0.054	0.052	0.009

**Annex Table 11.A.14. Average omission rate by assessment hour in the CBA for all domains**

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Math Linear - trend*	0.08	0.098	0.017
Math Linear - new*	0.067	0.074	0.007
Math MSAT - trend*	0.071	0.09	0.019

Math MSAT - new*	0.048	0.058	0.01
Reading Core Items	0.036	0.052	0.015
Reading Stage 1 and 2	0.063	0.078	0.015
Creative Thinking	0.041	0.053	0.012

**Annex Table 11.A.15. Average proportion correct (P+) by cluster position in new PBA**

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Reading	0.623	0.614	0.601	0.583	-0.040
Science	0.480	0.481	0.468	0.462	-0.018
Mathematics	0.387	0.385	0.373	0.356	-0.029

**Annex Table 11.A.16. Average proportion of omitted responses by cluster position in new PBA**

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Reading	0.046	0.049	0.055	0.067	0.022
Science	0.061	0.055	0.067	0.074	0.013
Mathematics	0.095	0.090	0.095	0.110	0.015

**Annex Table 11.A.17. Example for use of plausible values for partitioning the error**

Plausible value	0-10 books at home		11-25 books at home		26-100 books at home		101-200 books at home		201-500 books at home		500+ books at home	
	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	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

**Annex Table 11.A.18. Panel Two of Example for use of plausible values for partitioning the error**

Estimate	429.41	474.29	512.91	538.83	560.30	546.79
Sampling Error	3.43	3.23	2.40	2.65	3.04	4.76
Measurement Error	0.42	0.87	0.43	1.42	1.02	0.85
Standard Error	3.46	3.34	2.44	3.00	3.21	4.83

**Annex Table 11.A.19. Number of trend (linking) items and new items by domain and mode of assessment**

	CBA Trend	CBA New	CBA Total	PBA	New PBA
Mathematics	74	159*	233	71	64
Reading	196*		196	87	66
Science	115		115	85	66
Reading Fluency	65		65		79
Financial Literacy	40*	5	45		
Creative Thinking		32*	32		

Note: \*Dropped items: CMA112Q02, CR547Q07S, DF082Q01C, and DT520Q01C, DT560Q01C, DT560Q02C, DT450Q01C, DT450Q02C and DT450Q03C

**Annex Table 11.A.20. Unweighted calibration sample size by domain and mode of assessment**

	CBA	PBA and New PBA
Mathematics	561,556	15,768
Reading	245,800	13,401
Science	245,715	13,209
Financial Literacy	42,068	
Creative Thinking	144,492	

**Annex Table 11.A.21. Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new mathematics items in the main survey**

MODEL	# of Parameters	AIC	BIC	Log Penalty	Improvement
Independence	NA	NA	NA	0.668	NA
Unidimensional	751	11861255	11869412	0.5794	99.8%
Two-dimensional	1002	11812371	11823248	0.5792	100.0%

Note: Log penalty (Gilula & 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.

**Annex Table 11.A.22. Distribution of the items to the mathematics subscales**

Content Scale			Process Scale		
Subscales	Trend	New	Subscales	Trend	New
Change and Relationships	17	38	Employing Mathematical Concepts, Facts and Procedures	24	51
Quantity	21	55	Formulating Situations Mathematically	11	37
Space and Shape	17	26	Interpreting, Applying and Evaluating Mathematical Outcomes	10	47
Uncertainty and Data	19	41	Reasoning	29	25
Total:	74	160	Total:	74	160

Note: CMA112Q02S (Content Scale - Quantity; Process Scale - Reasoning) was included in the counts above but was ultimately dropped during scaling for all countries.

# 12 Data Management Procedures

## Introduction

In PISA, as in any international survey, standards and requirements for data collection guide the creation of an international database that allows for valid within-and-cross-country comparisons and inferences to be made. For both paper-based assessments (PBA) and computer-based assessments (CBA), these standards and requirements are developed with three major goals in mind: consistency, precision, and generalisability. To support these goals, data collection and management procedures are applied in a common and consistent way across all data to ensure data quality. As such, “data management” within the scope of the PISA survey refers to a collective set of procedures and tasks that each country performs to produce a verified, national database. With these procedures, national teams can 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 were conducted with the national representatives of participating countries to reflect country expectations in agreement with PISA 2022 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 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. Such deviations, at times, may compromise data quality and/or render data corrupt, or unusable. For example, it may be the case that implementing non-standard testing procedures might, in turn, 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 of large-scale assessment administration and the compressed PISA timeline, 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 are generalised to suit the individual cognitive test instruments and background questionnaire instruments used in each participating country. As a result, a suite of products is provided to countries to assist national teams in handling data management tasks in a standard way to prepare the national database and minimise errors. These products include a comprehensive data management manual, training sessions, as well as a range of other materials, including the data management software.

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 by ETS Data Management to ensure data quality:

- Developed standards, guidelines, and recommendations for data management.
- Provided national teams with the data management software and developed data management manuals for modes of administration (PBA and CBA) as well as customized codebooks to support proper data capture.
- Facilitated data trainings and webinars and created hands-on, training resources (e.g., training exercises, lessons, and resource guides) for guided practice in building the national database and verifying data.
- Provided high-touch support for national team queries throughout the data management lifecycle.
- Enhanced data quality and verification procedures considering new context or situations during processing and cleaning data the international and national level.
- Prepared databases and reports for use by contractors, OECD, and the National Centres.
- Prepared interim and final data products (e.g., Data Explorer, compendia files) for dissemination to National Centres, the OECD, and, eventually, the public.

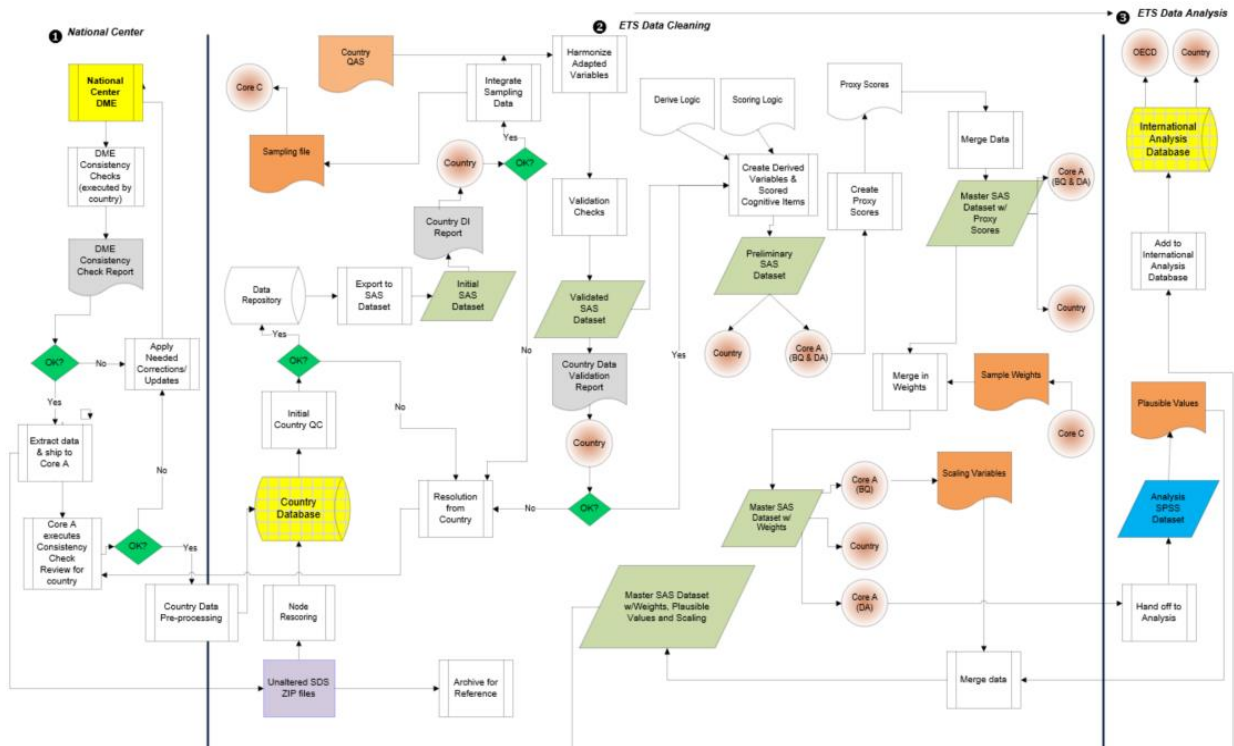
Ensuring compliance with technical standards also involved close collaboration with project partners. In PISA 2022, ETS Data Management worked closely with the all consortium members to ensure all data capture and quality procedures were accurately executed.

### ***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 (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 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 tasks at the national level. In addition, 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 data related 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 (e.g. Student Delivery System, Open-Ended Coding System); manual capture of data after scoring for paper-based instruments; export/import of data required for coding (e.g. occupational coding); and data verification and validation with a series of consistency and validity checks.

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 (DME). Figure 12.1 provides the workflow of the data management process for PISA 2022.

Figure 12.1. Overview 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 2022 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 school administrator responses on a computer platform into electronic data files provided a challenge and an opportunity for the accurate transcription of those responses as well as the collection of the associated process data, such as types of response actions and timing of those actions. It also requires a system that can accept and process these electronic data and their variety of formats as well as supports the manual entry of data from paper booklets and forms. To meet this challenge, ETS acquired a license for the use of the Data Management Expert (DME) software, which had previously proved successful in the collection and management of the data for the PISA 2015, PISA 2018, and PISA for Development large-scale surveys as well as the survey for adult skills (PIAAC) under a separate contract.

The DME 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, and does not require an internet connection to operate. It operates on a separate database file using SQLite constructed according to strict structural and relational specifications that define the data codebook. This codebook is a complete catalogue of all the data variables to be collected and managed, which are then arranged into well-defined datasets that correspond to the various instruments involved in the administration of the assessment. Before the datasets are created and ready for input processing, the application first validates the structure of the codebook to ensure the integrity of the database.

The first step in the data management process is to identify the different electronic and paper instruments, booklets, and forms that are to be collected and managed within each national centre and determine the variables to be collected from each instrument. These instruments and forms are then mapped into datasets, each containing their appropriate variables to form the international codebook, which will be the basis for every national codebook, whether the country is conducting the assessment on paper or computer. The international codebook is thoroughly checked, verified, and tested using marked up paper instruments as well as electronic data files that were created during testing of the various platforms.

The next step is the generation and testing of the national codebooks. Many of the variables used in the assessment and codebooks for PISA follow a systematic naming convention that provides additional information to the user. Annex Table 12.A.2. describes the naming convention used in the codebooks and analysis.

Each national codebook is a copy of the international codebook where the datasets corresponding to national options implemented in the country are shown and the rest are hidden. For example, all codebooks for PBA countries will have the datasets corresponding to CBA instruments hidden from view and operation. In addition, the codebooks for CBA countries will have all adapted and national questions that were coded into the Questionnaire Adaptation Tool (QAT, described in Chapter 7) added to the appropriate datasets. The CBA codebooks are also tested using available test data obtained from the country's student delivery platform and the online questionnaire system. PBA countries, as well as CBA countries with the paper-based Parent Questionnaire, are given the option of providing national translations of all items in the paper instruments to be included in their national codebooks.

The codebook is delivered to each country as a national "template" file, containing the metadata the DME application uses to build the database file. The NDM must confirm that the template file will create an accurate codebook that supports the appropriate datasets for their national options. To verify nationally adapted variables and/or added national variables, CBA countries are then requested to also import available test data to confirm proper data capture. For PBA countries, variables must be added and adapted first to the questionnaire datasets, as there are no online QAT questionnaire data available for these countries. They are then required to test these adaptations and added variables with the manual entry of the questionnaire data to confirm that the variables are properly configured, in their correct sequence, and with their correct translations, when applicable. Similarly, CBA countries with the Parent Questionnaire option, a paper-based option, must also add and test their national adaptations to the corresponding dataset. After making all necessary modifications to and testing of their national codebook, every country is requested to send a copy of the codebook to Data Management so it can be reviewed for consistency and use in the Main Survey.

The DME application permits three levels of password-controlled access to the database – Administrator, Manager, and User. The Administrator level has complete access to all the database operations as well as the data tables and codebook-related tables. This level is reserved for Data Management. The Manager level is designated for the NDM in each country and includes the ability to make changes to the codebook, create and delete data tables and create User accounts and passwords, among other capabilities. The User level is assigned by the Manager for the purpose of creating clones of the project Master database to be used for manual data entry on multiple platforms. The DME application is designed to work in a distributed environment so that these individual clone databases can be easily merged into the master database.

For the PISA survey, there are three, recommended modes for input of data into the DME application: manual data entry, import from Excel or CSV file, and special import of extracted data from student delivery, sampling, and coding systems.

Manual data entry provides for the direct entry of data values into a targeted dataset through an interface that presents the description, format, and valid codes of each data element to be entered and validates each entered value. The type of forms that can be entered vary from a simple linear form, such as a



questionnaire, to a series of booklets or forms that each contain a prescribed sequence of blocks of item data, such as the cognitive booklets. The entry of the booklet/form number determines which variables are to be presented for entry and in what order. The manual entry mode is used primarily by PBA countries as well as those CBA countries when using the Parent Questionnaire option.

If a PBA country has its own data entry procedures in place, the data from these processes can be directly imported from Excel or CSV files where the first row/record contains the names of the variables whose data are in the corresponding columns. Again, all input data values are validated against the codebook and if any unexpected or out of range data values are found, the process stops. This import process has a corresponding export process to create files, typically Excel and CSV, from designated datasets. The two processes can be effectively used to move data into and out of the database. The export process for CSV files also produces syntax files for reading the exported data into SPSS or SAS so that separate analyses of the data can be performed with those applications.

The Export and Import functions also include options for exporting and importing data for occupational coding. When the Export/Import for occupational coding menu items are chosen, data will be exported from/imported into multiple datasets. The resulting files will be a “pair” of macro-enabled Excel files for each questionnaire language code found in the database, one primary file and a second identical copy of the file to be used for double coding. When national teams complete the occupation coding and verify double coding agreement (through the internal check macro within the file) only the primary coded file is imported into the database.

The PISA Imports menu option contains specialized procedures designed to extract data from files delivered by the various electronic sources: the student delivery system (SDS), the online school and teacher questionnaires, the open-ended coding system (OECS), and ACER Maple sample management system. The DME application creates a log file for each imported data file to record the action for each data element encountered. All invalid data values are replaced with designated missing values and a record of that activity is added to an internal log table within the database.

It is the Data Manager’s responsibility to schedule and coordinate the various activities associated with the collection, entry and validation of the data in the database. They are typically allowed eight weeks after the last administration of the survey to gather and integrate the collected data into the database, including time for the human scoring of the cognitive items, and to perform all checks on the integrity and consistency of the data. For this last task the DME application provides the ability to perform various checks on the database. Two of them, the validation check and the Unique ID check, rarely yield actionable results as all methods of integrating data into the database undergo a validation check at the point of entry, and each dataset is designed so that duplicate ID’s can also be detected and prevented from entry into the database.

The Record Consistency check is a series of individual reports that are designed and scripted by Data Management to assist national teams with verifying:

1. Consistency between the absence codes in the sampling dataset and each of the other student datasets to determine if a student marked as absent has data in a related dataset or vice versa.
2. Consistency between the student demographics in the sampling dataset and the Student Background Questionnaire dataset.
3. Consistency between the cognitive response data files and their corresponding OECS datasets to ensure that all respondents received codes for the open-ended items.
4. Consistency between the questionnaire datasets and the cognitive datasets (i.e., whether a student took both sessions of the assessment).
5. Data entry inconsistencies of paper-based instruments.
6. Identification of missing response or coded occupational data.
7. Counts of certain aspects of the database, such as number of students by language of survey.

8. Consistency for the School and Teacher datasets related to participation, questionnaire data, and sampling information.
9. Identify the contents of specific inner tables, such as the “ImportValueErrors”, which captured all conversions of invalid data values into missing values.

These reports can be downloaded from the application to an Excel file. The NDM must review all of the cases identified in each report and, for all cases except the cases flagged in the “ImportValueErrors” check, the NDM should resolve the noted discrepancies or provide an explanation for why they could not be resolved. In addition to the Double-key entry report in the Record Consistency check, which checks for mis-matched IDs across datasets, there is also a separate Double-Key Data Entry check in the DME that is to be executed for all paper instruments, including the Parent Questionnaire. The **Double-Key Data Entry check** identifies inconsistent data values entered across corresponding data sets, such as, SBP1 and SBP2, the datasets containing the student questionnaires as entered by Key Entry Operator 1 and Key Entry Operator 2. For this check, the NDM must resolve all discrepancies before proceeding to the next step.

When the NDM is satisfied that all data that could be collected has been properly placed in the database and all discrepancies have been resolved or explained, the DME provides an export function that will create a read-only copy of the database where any variables that are designated for suppression (e.g., Personally Identifiable Information) are set to null values. This export database, along with the annotated consistency report document and, for CBA countries, a set of zip files containing all the electronic files that were imported into the database, are submitted to Data Management via a secure FTP site.

### ***Pre-processing – National Database and Corresponding Files***

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.

Data submission from countries included any “unprocessed” files, or files that the DME software was not able to import. Data Management made great efforts to recover as much of this data as possible by repairing the files or finding and importing into the database a usable version from the PISA Uploads Server. To specifically handle the unique cases observed in PISA 2022, an additional file recovery tool was developed to expedite data recovery.

Running the DME software’s Record Consistency Checks outlined above was one of the first quality control 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 the quality of the data received.

All sampling data (variables and values) was verified against approved sampling data from the sampling international contractor, Westat, at the student-level and, if applicable, at the teacher-level as well.

These checks, in addition to other internal checks for coding, missing data, and student/teacher tracking data alignment with approved sampling forms, were executed upon receipt of the data. Reported inconsistencies returned from these checks were compiled and sent 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 documented 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 round of consistency checks to be sure all issues were resolved and/or documented. This initial step of processing (i.e., returning data inconsistencies to the National Centres and receiving a revised database) was an iterative process of data

review and validation. Once issues were resolved or documented, the data continued to the next phase of the internal process – loading the database into the cleaning and verification software.

## **Data Processing and Cleaning System**

### *Loading the SQLite database into the Processing and Cleaning System*

With all 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 an efficient.NET application that uses SQL and SAS to merge and process datasets.

During the processing phase, one or two analysts independently loaded each national databases into the processing software, focusing on one country at a time, to complete all necessary phases of quality assurance. Once complete, SAS and SPSS datasets were delivered to the country, and other contractors for review and analyses.

The first step in this process was to load the pre-processed national database, an SQLite database, into the ETS Data Management cleaning and verification software. 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 2022, patch files were released by ETS Data Management and applied to the SQLite database by the National Data Manager to address issues in the codebook for proper data capture in the DME software. For example, a patch may be issued if an item was misclassified as having 4 response options instead of 5.
- The country database had the correct profile as dictated by the international options (e.g., Financial Literacy, *Une Heure* form, etc.) selected by the country.
- The number of cases in the data files by country/language agreed 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.

## **Cognitive Assessment Data Processing**

### *Integration*

After the initial load of data and completion of early processing checks, the database entered the next phase of processing: Integration. 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 for each of the respondent types: student, school, and teacher (where applicable). Additionally, Parent questionnaire data was merged with their child/student data.

During data processing, the integration phase was critical because the Data Management contractor was able to analyse the data collected within the context of the sampling information supplied by the sampling contractor. Using this sampling information –captured in the Student Data File and Teacher Data File – extensive quality control checks were applied to the data in this phase. Over 100 quality assurance checks were performed on the database. As a result of these quality assurance checks, a data quality report was generated and delivered to countries to resolve outstanding issues and inconsistencies. This report was known as the Data Integrity (“DI”) Report.

In this report, the Data Management contractor 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):

- Cognitive test (FORMCODE) variable was blank or not valid.
- Student was missing key data needed for sampling and processing.
- Student was not within the allowable age for the assessment.
- Student was not represented in the Sampling Data (Student Data File).
- Students was marked absent yet had a response record.
- Student’s grade was lower than allowed.
- Student’s assessment path misaligned with the multi-stage, adaptive design.
- A teacher was marked as a “non-participant,” yet response data existed for that teacher.
- The DI report was packaged along with a series of other quality control reports (i.e., harmonisation report and validation report, see “Background Questionnaire Assessment Data Processing”) for national team review. When reviewing the report, National Centre teams were asked to review flagged inconsistencies from the report and correct data issues in the national database. National teams were instructed to complete the report review and revision of data within a specific timeframe for resubmission to the Data Management contractor. Additionally, national teams documented all data revisions in the DI report and returned the report to the Data Management contractor for review.
- After receiving the revised database and all documentation, the Data Management contractor repeated the pre-processing phase to ensure no new errors were reported and, if no issues or errors were found, the Data Management analyst re-executed the Integration step. As with the pre-processing consistency checks phase, the Integration step might have 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 to resolve issues.

In addition to quality assurance reporting, a series of important data processing steps occurred during the Integration phase:

- **Item Cluster Analysis:** For the purposes of data processing, it is often convenient to be able to disaggregate a single variable into a collection of variables. To this end, a respondent’s single booklet number was generated 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 to detect any inconsistencies between information in the student delivery system and information in 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 captured accurately. This was not necessarily true in the case of computer-based administrations. While the student delivery system captures a student's response, it does not 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 are merged to the country database.
- **Process Data:** The student delivery system also produced log files where process data could be extracted for further analysis. Process data including the total response time, response time to first action, number of visits, number of short visits, and the number of actions were extracted by specialized tools and then verified by the Data Management contractor through a series of quality control checks. Such quality control checks identified inconsistencies or situations of unreasonableness (e.g., duplicated records, out-of-range values, system or operational issues, total unit duration is higher than item time). Once inconsistent results were either resolved or explained, the data were provided to psychometric teams for further analysis.
- **SDS Post-processing:** Necessary changes in the student delivery system were sometimes detected after the platform was already in use. For example, a test item that was scored by the delivery system may have had an error in the interpretation of a correct response, which was corrected in post-processing. These and other issues were resolved by the delivery system's developers and new scored response data was processed, issued, and merged by the Data Management contractor.
- **Multi-Stage Adaptive Testing:** For both the Reading and Mathematics CBA, counts and percentages were produced for each country. Such counts identified the breakdown of each stage by performance to confirm that the student delivery platform's routing worked as expected during the assessment.

### *Scoring*

After initial integration of the data, the next phase of data management processing involved parallel processes that occur with assessment data:

- Scoring of test responses captured in paper booklets.
- Treatment of CBA human-coded items.
- Additional checks of cognitive items.

### *Scoring overview*

The goal of the PISA assessment is to ensure comparability of the assessment results across countries. As a result, scoring of the responses to the test items 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. The Data Management contractor 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 each response. The Data Management contractor was not only responsible for generating the syntax to implement the scoring rules 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 variables where data was expected, the Data Management contractor 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 mis-coded 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/processing were completed. If the reported data inconsistencies were resolved, the scoring process was deemed complete, and the data proceeded to the next phase of processing.

The scoring variables also served as a valuable data quality check. If any items appeared to function unexpectedly (i.e., too difficult, too easy, or unusually high missing rates), further investigation was carried out to determine if a booklet printing or translation error occurred or if systematic errors were introduced during the administration, data load, or data entry.

Once the Integration and Scoring steps were complete, the next phase of data cleaning involved the validation of the background questionnaire data, i.e., harmonisation of national adaptations and verification of questionnaire response data.

## **Background Questionnaire Assessment Data Processing**

### *Harmonisation*

#### **Harmonisation, or harmonised variables**

As mentioned earlier, 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 or contexts. These adaptations are referred to as “national adaptations.” While able to capture country contexts, these adapted variables needed to be mapped into the corresponding international variable for cross-country comparison.

More specifically, 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 using SAS code. This code was implemented into the cleaning system to handle these national variables during processing.

Additionally, harmonisation consisted of mapping 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 (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); or perhaps two questions were merged.

#### **Overview of the workflow**

To capture the appropriate adaptation and harmonisation, changes to variables by national teams were proposed during the translation and adaptation process. National adaptations for questionnaire variables were agreed upon by the Background Questionnaire contractor. These discussions regarding adaptations happened in the negotiation phase between the country and the contractor as well as the translation verification contractor – prior to data submission to ETS. 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 corresponding harmonisation mapping. 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 Background Questionnaire contractor, with the assistance of the translation verification contractor, to ensure the QAS was complete and reflected the country's intent and interpretation.

Issues surrounding national adaptations and/or the harmonisation code produced by the cleaning software, often, involved consultation with the national team as well as the Background Questionnaire contractor. Both the Background Questionnaire contractor and the national team were responsible for reviewing the harmonisation report produced by the Data Management contractor during processing to verify national adaptations and corresponding mappings. Requested updates or changes were documented in the harmonisation report, the country QAS file, and the cleaning system harmonisation code. As a result of updates, a new harmonisation report was generated and delivered to the national team and the Background Questionnaire contractor for final review and approval.

### *Validation*

After the Harmonisation step, 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, the Data Management contractor collaborated with the Background Questionnaire contractor to develop a series of validation checks that were performed on the data following harmonisation.

Validation checks are a set of consistency checks that provide National Centres with more detail concerning extreme and/or inconsistent values in their data. Issues detected by these checks were displayed in a validation report, which was shared with countries and contractors to observe these inconsistencies and potentially make improvements for the next cycle of PISA. Consistent with PISA 2018, national teams did not make changes to revise these extreme and/or inconsistent values in the report. Rather, national teams were instructed to leave the data as-is and make recommendations for addressing these issues in the data collection process during the next phase from Field Trial to Main Survey, or the next cycle of PISA.

Generally, validation checks captured inconsistent student, school, and teacher data. For example, these checks captured an inconsistency between the total number of years teaching, and the number of years teaching at a particular school (TC00701); 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 (ST059Q02). Throughout the PISA cycle, these validation checks often served as valuable feedback to check on the data quality.

#### **Treatment of inconsistent and extreme values in PISA 2022 main survey data**

Following the approach implemented in PISA 2015 and PISA 2018 for extreme and/or inconsistent values within national data, the Data Management contractor, the Background Questionnaire contractor, and the OECD agreed on the implementation of specific range restriction rules applied during data cleaning that would manage extreme and/or inconsistent values. These values would be invalidated across all country databases.

Building on the range restriction rules developed in PISA 2015 and used in PISA 2018, the following principles were observed in the special handling of these inconsistent and/or extreme values:

- 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; or, if an inconsistency existed between age and seniority, the proposed rule may invalidate seniority, but keep “age.”).
- Apply stringent consistency and validity checks while computing derived variables. With this principle, the original values may be kept, while the values for the derived variable may have applied an “invalid” rule.

The specific range restriction rules for PISA 2022 are presented in Annex Table 12.A.3.

### *Derived variables*

Code in SAS to create derived variables was generated by the Background Questionnaire contractor for implementation into the cleaning system at this step in the process. The code to create derived variables 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 that outlined the derived variables that were calculated from PISA MS data.

As further explained in the MS Analysis Plan, for all questions in the MS questionnaires, regardless of whether they served as a basis for derived variables or not, the international database contains item-level data as obtained from the delivery platform. For any derived variables, whenever possible, these were specified consistent with previous cycles of PISA. In terms of this alignment, the first choice was alignment with PISA 2012, to enable comparison on math-related variables. The second choice was alignment with PISA 2018. This aimed to strike a balance and stability across recent and future cycles. A list of PISA 2022 Main Survey non-item response theory (IRT) derived variables (“simple indices”) is available in Annex Table 12.A.5.

As part of quality control, all derivations were verified by the Background Questionnaire contractor. Any updates or recoding made to the derived variable code were completed, documented, and redelivered to the Data Management contractor for use in the cleaning system. Data files were refreshed to implement any changes to the code or the variables.

### ***Deliverables***

After all data processing steps were complete and all updates to the data were made by national teams to resolve any issues or inconsistencies, the final phase of data processing included the creation of deliverable files for specific contractors (e.g., Westat for Sampling, or ETS for Data Analysis) 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 process contained critical additions to the data – such as the addition of proxy scores, plausible values, background questionnaire scales, and sampling weights (student and teacher). The dynamic feature of the cleaning system allowed for the Data Management contractor to generate customized files for delivery at specific phases of the project lifecycle.

To produce these customized files for specific clients at specific phases of the project, 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

To prepare for the public release of PISA 2022 main survey data, the Data Management contractor 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 preliminary sampling weights and proxy proficiency scores for analysis. These data were later updated to include final sampling weights, 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.

The following data files were delivered:

- The Student combined data file contained all student responses to the test items (raw and scored), background questionnaire items, and optional questionnaire items such as Parent Questionnaire, Well-Being (WB) Questionnaire, Information and Computer Technology Literacy Familiarity (ICT) Questionnaire. These files included all raw variables, questionnaire indices, student weights, replicate weights, and plausible values.
- The School data file contained all response data collected with the School Questionnaires. These files included all raw responses, school-level base weights, questionnaire indices, and other derived variables.
- The Teacher data file contained response data from the Teacher Questionnaire. These files included all raw responses, questionnaire indices, derived variables, and teacher weights.
- The Financial literacy data file contained response data from the financial literacy cognitive and background questionnaire items. These files included all raw variables, questionnaire indices, sampling weights, replicate weights, and plausible values.
- The Masked international database, which combined the data from all participating countries. To preserve country anonymity in this file, key identifying variables were masked following specific guidelines from the OECD Secretariat that included issuing alternate codes or required special handling for country identifiers.
- The preliminary, national version of the Public Use File (PUF) was produced toward the end of the PISA 2022 main survey and provided the National Centre with the opportunity to review their data before the final public release. These data included all country-requested variable suppressions. More information about country-level variable suppressions is included in Annex Table 12.A.3.

In addition to these data files, a series of analysis reports were produced by the Data Analysis team and delivered by the Data Management contractor to National Centres for quality control, data validation, and further national analyses. These reports were also used to evaluate the plausibility of the distributions of background characteristics and the performance results by subgroups, especially evaluating the extent to which they agree with expectations based on external or historical information. These reports included:

- BQ Crosstabs: A report containing frequencies of numerical, categorical variables from the country's Background Questionnaire (BQ). To aide countries in reviewing their BQ variables for potential translation or coding errors, flagging for outliers as compared across countries were included in this report.
- BQ MSIGS: A report containing summary statistics for all numerical variables from the country's Background Questionnaire.
- BQ SDTs: A set of reports 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.

- Codebook Descriptives Report: A report that includes 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 and includes descriptive statistics for all variables.
- Cognitive Summary Analysis Reports: A comprehensive report that included a series of key statistics and flags across item analysis (IA), coding reliability, and item response theory (IRT) reports to identify items that, based on the empirical data, are most likely to require careful review and feedback by national teams.
- Item Analysis Reports: A set of reports that provided summary information about the response types given by the respondents to the cognitive items. They contained, for each country, various statistics (e.g., count, percent, mean cluster score) of students choosing each option for multiple-choice items or the percent of individuals receiving each score in the scoring guide for the constructed-response items.

### ***The Public Use File - Included Records***

When preparing for the final public use file (PUF), the following records were included in the database:

#### *Student files*

- Includes one records per respondent<sup>1</sup> that met the international target population definition and that passed validation, adjudication, and weighting.

#### *School files*

- Includes one record per participating school – specifically, one record for any school with a student included in the PISA sample regardless of whether the school returned the School Questionnaire.

#### *Teacher files*

- Include one record for each teacher that met the international target population definition and that passed validation, adjudication, and weighting<sup>2</sup>.

#### *Financial literacy student files*

- One record per student respondent that met the international target population definition and that passed validation, adjudication, and weighting; and that responded to a cognitive form that included Financial Literacy items (Forms 67 – 74), or included Mathematics and reading items (Forms 1-12).

### ***Categorising missing data***

Within the data files, the coding of the data distinguishes between six different types of missing data:

1. System Missing/Blank – used to indicate that the respondent was not presented the question according to the survey design or ended the questionnaire early, or data loss.
2. No Response – used to indicate the respondent had an opportunity to answer the question but did not respond. For derived variables, it is often used as an indicator for all different types of missing data.
3. Invalid – used to indicate that the response was not appropriate or contradicted a prior response, e.g., the response to a question asking for a percentage was greater than 100.

4. Not Applicable – used to indicate in the questionnaire that the question was not asked by design or could not be determined due to a printing problem or torn booklet, or due to within-construct matrix sampling design. In the cognitive data, it is used to indicate that the question was dropped/deleted during item calibration and not used during scaling.
5. Valid Skip – used in the questionnaire data to indicate that the question was not answered because a response to an earlier question directed the respondent to skip the question.
6. Not Reached – used in the cognitive scored variables to indicate that a student was unlikely to have seen the question and the response should be treated as such.

### ***Data management and confidentiality, variable suppressions***

During the PISA 2022 cycle, some country regulations and laws restricted the sharing of certain data 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 could impact the analytical utility of the data. Therefore, both goals must be carefully balanced. As a general directive for PISA 2022, 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 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. To suppress data for the public use files, an invalid code was applied during the final step of data file creation in the cleaning system<sup>3</sup>. A listing of suppressions at the country variable-level is in Annex Table 12.A.4.

### **Notes**

- 
1. To be considered a “respondent” the student must have responded to at least half of the number of test items in his or her booklet/form; or at least one test item response and a minimum number of responses to the student background questionnaire.
  2. Teachers who were absent, excluded, or refused to participate in the session may be marked as a “non-participant.”
  3. PISA national participants also had the opportunity to request a withdrawal of data. These requests were managed by the OECD and implemented by the Data Management contractor. The withdrawal of data involves removing data (e.g., records from specific regions) from data files and reports (including public-use files) for country-specific reasons. The request to withdrawal data required thorough discussion with the OECD and approval.

## Annex 12.A. Additional Data Management Items

Annex Table 12.A.1. Chapter 12 PISA Variable Naming, Codes, and Suppressions

Table	Title
Web Table 12.A.5	PISA Non-IRT Derived Variables Code

StatLink  <https://stat.link/8piamj>

Annex Table 12.A.2. PISA Variable Naming Convention

First Character	Second Character	Next Three Characters	Next Three Characters	Last Character
Indicates whether the variable is derived from the paper- or computer-based assessment	Indicates the cognitive domain for the related item	Is a unique numeric item identifier within each domain	Include of a "Q" and a two-digit numeric item part code.	Indicates additional information of the type of information captured.
<p><b>P</b> for <b>paper-based items</b> (Note: some of the paper-based reading and science trend items do not have "P" as the first character and, instead, may begin with "R" or "S" – see "Second Character" column)</p> <p><b>C</b> for <b>computer-based items</b> (Note: Creative Thinking items do not have "C" as the first character, these variables begin with "T" – see "Second Character" column)</p> <p><b>D</b> for <b>computer-based, human-coded items</b></p>	<p><b>M</b> for <b>Mathematics trend items</b> <b>MA</b> for <b>Mathematics new items</b> <b>R</b> for <b>Reading trend items</b> <b>S</b> for <b>Science trend items</b> <b>F</b> for <b>Financial Literacy items</b> <b>T</b> for <b>Creative Thinking items</b></p>			<p><b>S, SA, SB, SC, etc.</b> for the <b>scored response</b> C for a human-coded computer-based code R, RA, RB, RC, etc. for the actual response TT for the total timing F for the time to first action A for the number of actions V for the number of visits VS for the number of short visits</p>

Annex Table 12.A.3. PISA 2022 Range Restrict Code

Sequence	Dataset STU, SCH, TCH)	Description Code	SAS Code
STUDENT			
1	STU	INVALIDATE IF NUMBER FOR AN INDIVIDUAL'S WEIGHT IS NEGATIVE.	IF ((WB151Q01HA < 30) OR (WB151Q01HA > 250)) AND (NOT MISSING(WB151Q01HA)) THEN WB151Q01HA=.;
2	STU	INVALIDATE IF NUMBER FOR AN INDIVIDUAL'S HEIGHT IS NEGATIVE.	IF ((WB152Q01HA < 90) OR (WB152Q01HA > 230)) AND (NOT MISSING(WB152Q01HA)) THEN WB152Q01HA=.;
3	STU	INVALIDATE IF NUMBER FOR AN INDIVIDUAL'S CLOSE FRIENDS IS MORE THAN 50. (LISTED IN MAT'S EMAIL FROM 4/8/19 BUT NOT IN THIS EXCEL FILE)	IF (WB156Q01HA > 50) THEN WB156Q01HA =.;

4	STU	INVALIDATE IF NUMBER OF CLASS PERIODS PER WEEK IN MATHEMATICS LESSONS (ST059Q01TA) IS NEGATIVE OR GREATER THAN 75	IF (ST059Q01TA > 75 OR ST059Q01TA < 0) AND NOT MISSING(ST059Q01TA) THEN ST059Q01TA =.I;
5	STU	INVALIDATE IF NUMBER OF TOTAL CLASS PERIODS IN A WEEK (ST059Q02JA) IS NEGATIVE OR GREATER THAN 120 OR LESS THAN 10.	IF (ST059Q02JA > 120 OR ST059Q02JA < 0) AND NOT MISSING(ST059Q02JA) THEN ST059Q02JA =.I;
6	STU	INVALIDATE IF A CHILD'S ISCED LEVEL EQUALS 2 AND SELECTS THAT HE OR SHE HAS REPEATED ISCED 3 ONCE OR MULTIPLE TIMES	IF INT(ISCEDP/100)=2 AND (ST127Q03TA=2 OR ST127Q03TA=3) THEN ST127Q03TA =.I;
SCHOOL			
1	SCH	INVALIDATE IF TOTAL NUMBER OF COMPUTERS (SC004Q02TA) IS NEGATIVE.	IF (SC004Q02TA < 0) AND NOT MISSING(SC004Q02TA) THEN SC004Q02TA =.I;
2	SCH	INVALIDATE IF TOTAL NUMBER OF COMPUTERS (SC004Q03TA) IS NEGATIVE.	IF (SC004Q03TA < 0) AND NOT MISSING(SC004Q03TA) THEN SC004Q03TA =.I;
3	SCH	INVALIDATE IF TOTAL NUMBER OF WHITEBOARDS (SC004Q05NA) IS NEGATIVE.	IF (SC004Q05NA < 0) AND NOT MISSING(SC004Q05NA) THEN SC004Q05NA =.I;
4	SCH	INVALIDATE IF TOTAL NUMBER OF DATA PROJECTORS (SC004Q06NA) IS NEGATIVE.	IF (SC004Q06NA < 0) AND NOT MISSING(SC004Q06NA) THEN SC004Q06NA =.I;
5	SCH	INVALIDATE IF TOTAL NUMBER OF COMPUTERS (SC004Q07NA) IS NEGATIVE.	IF (SC004Q07NA < 0) AND NOT MISSING(SC004Q07NA) THEN SC004Q07NA =.I;
6	SCH	INVALIDATE IF TOTAL NUMBER OF TABLETS OR E-BOOK READERS (SC004Q08JA) IS NEGATIVE.	IF (SC004Q08JA < 0) AND NOT MISSING(SC004Q08JA) THEN SC004Q08JA =.I;
7	SCH	INVALIDATE IF NUMBER OF DESKTOP OR LAPTOP COMPUTERS CONNECTED TO THE INTERNET (SC004Q03TA) IS GREATER THAN THE NUMBER OF DESKTOP OF LAPTOP COMPUTERS AVAILABLE TO STUDENTS (SC004Q02TA).	IF SC004Q03TA > SC004Q02TA AND NOT MISSING(SC004Q02TA) THEN SC004Q03TA =.I;
8	SCH	INVALIDATE IF TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01) IS NEGATIVE.	IF (SC018Q01TA01 < 0) AND NOT MISSING(SC018Q01TA01) THEN SC018Q01TA01 =.I;
9	SCH	INVALIDATE IF NUMBER OF FULL-TIME CERTIFIED TEACHERS (SC018Q02TA01) IS NEGATIVE	IF (SC018Q01TA02 < 0) AND NOT MISSING(SC018Q01TA02) THEN SC018Q01TA02 =.I;
10	SCH	INVALIDATE IF NUMBER OF FULL-TIME CERTIFIED TEACHERS (SC018Q02TA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q02TA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q02TA01 =.I;
11	SCH	INVALIDATE IF NUMBER OF FULL-TIME BACHELOR DEGREE TEACHERS (SC018Q08JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q08JA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q08JA01 =.I;
12	SCH	INVALIDATE IF NUMBER OF FULL-TIME MASTER'S DEGREE TEACHERS (SC018Q09JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q09JA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q09JA01 =.I;
13	SCH	INVALIDATE IF NUMBER OF FULL-TIME DOCTORAL DEGREE TEACHERS (SC018Q10JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC018Q01TA01).	IF SC018Q10JA01 > SC018Q01TA01 AND NOT MISSING(SC018Q01TA01) THEN SC018Q10JA01 =.I;
14	SCH	INVALIDATE IF NUMBER OF PART TIME CERTIFIED TEACHERS (SC018Q02TA02)	IF SC018Q02TA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q02TA02 =.I;

		EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	
15	SCH	INVALIDATE IF NUMBER OF PART TIME BACHELOR DEGREE TEACHERS (SC018Q08JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	IF SC018Q08JA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q08JA02 =.I;
16	SCH	INVALIDATE IF NUMBER OF PART TIME MASTER'S DEGREE TEACHERS (SC018Q09JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	IF SC018Q09JA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q09JA02 =.I;
17	SCH	INVALIDATE IF NUMBER OF PART TIME DOCTORAL DEGREE TEACHERS (SC018Q10JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC018Q01TA02).	IF SC018Q10JA02 > SC018Q01TA02 AND NOT MISSING(SC018Q01TA02) THEN SC018Q10JA02 =.I;
18	SCH	INVALIDATE IF TOTAL NUMBER OF FULL-TIME MATHEMATICS TEACHERS (SC182Q01WA01) IS NEGATIVE.	IF (SC182Q01WA01 < 0) AND NOT MISSING(SC182Q01WA01) THEN SC182Q01WA01 =.I;
19	SCH	INVALIDATE IF NUMBER OF FULL-TIME CERTIFIED MATHEMATICS TEACHERS (SC182Q06WA01) EXCEEDS TOTAL NUMBER OF FULL-TIME MATHEMATICS TEACHERS (SC182Q01WA01).	IF SC182Q06WA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q06WA01 =.I;
20	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS BACHELOR DEGREE TEACHERS (SC182Q07JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q07JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q07JA01 =.I;
21	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND MATH MAJOR (SC182Q08JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q08JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q08JA01 =.I;
22	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND PEDGOGY QUALIFCATION (SC182Q09JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q09JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q09JA01 =.I;
23	SCH	INVALIDATE IF NUMBER OF FULL-TIME MATHEMATICS ISCED 5 TEACHERS (SC182Q10JA01) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA01).	IF SC182Q10JA01 > SC182Q01WA01 AND NOT MISSING(SC182Q01WA01) THEN SC182Q10JA01 =.I;
24	SCH	INVALIDATE IF TOTAL NUMBER OF PART TIME MATHEMATICS TEACHERS (SC182Q01WA02) IS NEGATIVE.	IF (SC182Q01WA02 < 0) AND NOT MISSING(SC182Q01WA02) THEN SC182Q01WA02 =.I;
25	SCH	INVALIDATE IF NUMBER OF PART TIME CERTIFIED TEACHERS (SC182Q06WA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC182Q01WA02).	IF SC182Q06WA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q06WA02 =.I;
26	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS BACHELOR DEGREE TEACHERS (SC182Q07JA02) EXCEEDS TOTAL NUMBER OF PART TIME TEACHERS (SC182Q01WA02).	IF SC182Q07JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q07JA02 =.I;
27	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND MATH MAJOR (SC182Q08JA02) EXCEEDS TOTAL	IF SC182Q08JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q08JA02 =.I;

		NUMBER OF FULL-TIME TEACHERS (SC182Q01WA02).	
28	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS TEACHERS WITH BACHELOR DEGREE AND PEDAGOGY QUALIFICATION (SC182Q09JA02) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA02).	IF SC182Q09JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q09JA02 =.I;
29	SCH	INVALIDATE IF NUMBER OF PART TIME MATHEMATICS ISCED 5 TEACHERS (SC182Q10JA02) EXCEEDS TOTAL NUMBER OF FULL-TIME TEACHERS (SC182Q01WA02).	IF SC182Q10JA02 > SC182Q01WA02 AND NOT MISSING(SC182Q01WA02) THEN SC182Q10JA02 =.I;
30	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 =.I;SC016Q02TA =.I;SC016Q03TA =.I;SC016Q04TA =.I;
31	SCH	INVALIDATE IF PERCENTAGE OF TEACHING STAFF (SC025Q01NA) IS GREATER THAN 100%.	IF SC025Q01NA>100 THEN SC025Q01NA =.I;
32	SCH	INVALIDATE IF PERCENTAGE OF MATHEMATICS TEACHER STAFF (SC025Q02NA) IS GREATER THAN 100%.	IF SC025Q02NA>100 THEN SC025Q02NA =.I;
33	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH <HERITAGE LANGUAGE> DIFFERENT THAN <TEST LANGUAGE> (SC211Q01JA) IS GREATER THAN 100%.	IF SC211Q01JA>100 THEN SC211Q01JA =.I;
34	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH SPECIAL LEARNING NEEDS (SC211Q02JA) IS GREATER THAN 100%.	IF SC211Q02JA>100 THEN SC211Q02JA =.I;
35	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS FROM DISADVANTAGED HOMES (SC211Q03JA) IS GREATER THAN 100%.	IF SC211Q03JA>100 THEN SC211Q03JA =.I;
36	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WHO ARE IMMIGRANTS (SC211Q04JA) IS GREATER THAN 100%.	IF SC211Q04JA>100 THEN SC211Q04JA =.I;
37	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WHOSE PARENTS ARE IMMIGRANTS (SC211Q05JA) IS GREATER THAN 100%.	IF SC211Q05JA>100 THEN SC211Q05JA =.I;
38	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WHO ARE REFUGEES (SC211Q06JA) IS GREATER THAN 100%.	IF SC211Q06JA>100 THEN SC211Q06JA =.I;
39	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT INITIATED DISCUSSION ON CHILD'S PROGRESS (SC064Q01TA) IS GREATER THAN 100%.	IF SC064Q01TA>100 THEN SC064Q01TA =.I;
40	SCH	INVALIDATE IF PERCENTAGE OF PARENTS WHERE TEACHER-INITIATED DISCUSSION ON CHILD'S PROGRESS (SC064Q02TA) IS GREATER THAN 100%.	IF SC064Q02TA>100 THEN SC064Q02TA =.I;
41	SCH	INVALIDATE IF PERCENTAGE OF PARENTS PARTICIPATED IN SCHOOL GOVERNMENT (SC064Q03TA) IS GREATER THAN 100%.	IF SC064Q03TA>100 THEN SC064Q03TA =.I;
42	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT VOLUNTEERED IN EXTRACURRICULAR ACTIVITIES (SC064Q04NA) IS GREATER THAN 100%.	IF SC064Q04NA>100 THEN SC064Q04NA =.I;

43	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT INITIATED DISCUSSION ON CHILD'S BEHAVIOR (SC064Q05WA) IS GREATER THAN 100%.	IF SC064Q05WA>100 THEN SC064Q05WA =.I;
44	SCH	INVALIDATE IF PERCENTAGE OF PARENTS WHERE TEACHER-INITIATED DISCUSSION ON CHILD'S BEHAVIOR (SC064Q06WA) IS GREATER THAN 100%.	IF SC064Q06WA>100 THEN SC064Q06WA =.I;
45	SCH	INVALIDATE IF PERCENTAGE OF PARENTS THAT ASSISTED IN FUNDRAISING (SC064Q07WA) IS GREATER THAN 100%.	IF SC064Q07WA>100 THEN SC064Q07WA =.I;
46	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 =.I; SC002Q02TA=.I; END;
47	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 =.I;
48	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH MARKS AT OR ABOVE (SC178Q01JA) AND BELOW PASSING (SC178Q02JA) IS GREATER THAN 100%.	IF SUM(SC178Q01JA + SC178Q02JA) >100 THEN DO; SC025Q01NA =.I; SC178Q01JA=.I; SC178Q02JA=.I; END;
49	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q01JA) IS NEGATIVE.	IF (SC168Q01JA < 0) AND NOT MISSING(SC168Q01JA) THEN SC168Q01JA =.I;
50	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q02JA) IS NEGATIVE.	IF (SC168Q02JA < 0) AND NOT MISSING(SC168Q02JA) THEN SC168Q02JA =.I;
51	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q03TA) IS NEGATIVE.	IF (SC168Q03JA < 0) AND NOT MISSING(SC168Q03JA) THEN SC168Q03JA =.I;
52	SCH	INVALIDATE IF TOTAL NUMBER OF NON-TEACHING STAFF (SC168Q04TA) IS NEGATIVE.	IF (SC168Q04JA < 0) AND NOT MISSING(SC168Q04JA) THEN SC168Q04JA =.I;
53	SCH	INVALIDATE IF TOTAL NUMBER OF FOREIGN LANGUAGES (SC174Q01JA) IS NEGATIVE.	IF (SC174Q01JA < 0) AND NOT MISSING(SC174Q01JA) THEN SC174Q01JA =.I;
54	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q01JA) IS NEGATIVE.	IF (SC213Q01JA < 0) AND NOT MISSING(SC213Q01JA) THEN SC213Q01JA =.I;
55	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q02JA) IS NEGATIVE.	IF (SC213Q02JA < 0) AND NOT MISSING(SC213Q02JA) THEN SC213Q02JA =.I;
56	SCH	INVALIDATE IF PERCENTAGE OF STUDENTS WITH MARKS AT OR ABOVE (SC178Q01JA) AND BELOW PASSING (SC178Q02JA) ARE BOTH ZERO.	IF (SC178Q01JA = 0 AND SC178Q02JA = 0) THEN DO; SC178Q01JA=.I; SC178Q02JA=.I; END;
57	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q01JA) IS NEGATIVE OR >1000.	IF (SC213Q01JA < 0) AND NOT MISSING(SC213Q01JA) THEN SC213Q01JA =.I; IF (SC213Q01JA >1000 THEN SC213Q01JA =.I;
58	SCH	INVALIDATE IF TOTAL NUMBER OF DAYS (SC213Q02JA) IS NEGATIVE OR >1000.	IF (SC213Q02JA < 0) AND NOT MISSING(SC213Q02JA) THEN SC213Q02JA =.I; IF (SC213Q02JA >1000 THEN SC213Q02JA =.I;
59	SCH	(SC175Q01JA, SC175Q02JA) THE MINUTES PER CLASS PERIOD SHOULD SET TO 1-120	IF (SC175Q01JA < 1) AND NOT MISSING(SC175Q01JA) THEN SC175Q01JA =.I; IF (SC175Q01JA >120 THEN SC175Q01JA =.I; IF (SC175Q02JA < 1) AND NOT MISSING(SC175Q02JA) THEN SC175Q02JA =.I; IF (SC175Q02JA >120 THEN SC175Q02JA =.I;

TEACHER



1	TCH	INVALIDATE IF NUMBER OF YEARS TEACHING AT SCHOOL (TC007Q01NA) EXCEEDS REPORTED AGE (TC002Q01NA) MINUS 15.	IF TC007Q01NA > (TC002Q01NA - 15) AND NOT MISSING(TC002Q01NA) THEN TC007Q01NA =.I;
2	TCH	INVALIDATE IF TOTAL NUMBER OF YEARS TEACHING (TC007Q02NA) EXCEEDS REPORTED AGE (TC002Q01NA) MINUS 15.	IF TC007Q02NA > (TC002Q01NA - 15) AND NOT MISSING(TC002Q01NA) THEN TC007Q02NA =.I;
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 AND NOT MISSING(TC007Q02NA) THEN TC007Q01NA =.I;
4	TCH	INVALIDATE IF SUM OF TEACHER EDUCATION OR TRAINING PROGRAMME OR OTHER PROFESSIONAL QUALIFICATION IS LESS THAN 98% OR GREATER THAN 102% (TC203Q01HA + TC203Q02HA + TC203Q03HA)	IF SUM( TC203Q01HA, TC203Q02HA, TC203Q03HA) > 102 OR SUM( TC203Q01HA, TC203Q02HA, TC203Q03HA) < 98 THEN DO; TC203Q01HA =.I; TC203Q02HA=.I; TC203Q03HA =.I;
5	TCH	INVALIDATE IF SUM OF TEACHER EDUCATION OR TRAINING PROGRAMME OR OTHER PROFESSIONAL QUALIFICATION DURING THE LAST 12 MONTHS IS LESS THAN 98% OR GREATER THAN 102% (TC204Q01HA + TC204Q02HA + TC204Q03HA)	IF SUM( TC204Q01HA, TC204Q02HA, TC204Q03HA) > 102 OR SUM( TC203Q01HA, TC203Q02HA, TC203Q03HA) < 98 THEN DO; TC204Q01HA =.I; TC204Q02HA=.I; TC204Q03HA =.I;
6	TCH	INVALIDATE IF NUMBER OF DAYS (TC257Q01JA) IS NEGATIVE.	IF (TC257Q01JA < 0) AND NOT MISSING(TC257Q01JA) THEN TC257Q01JA =.I;

### Annex Table 12.A.4. PISA Country Variable Suppressions

Country Variable Suppression	
<b>Austria</b>	
GRADE	SC016Q01TA
OCOD1 (2-digit)	SC016Q02TA
OCOD2 (2-digit)	SC016Q03TA
PROGN	SC016Q04TA
SC001Q01TA (recoding)	SCHLTYPE
SC002Q01TA	SCHSIZE (recoding)
SC002Q02TA	ST001D01T (recoding)
SC004Q01TA	STRATUM
SC014Q01TA	
<b>Belgium (French/German)</b>	
ST003D02T	
<b>Canada</b>	
CLSIZE	SC176Q01JA
MCLSIZE	SC182Q01WA01
SC002Q01TA	SC182Q01WA02
SC002Q02TA	SC182Q06WA01
SC003Q01TA	SC182Q06WA02
SC004Q01TA	SC182Q07JA01
SC018Q01TA01	SC182Q07JA02
SC018Q01TA02	SC182Q08JA01
SC018Q02TA01	SC182Q08JA02
SC018Q02TA02	SC182Q09JA01
SC018Q08JA01	SC182Q09JA02
SC018Q08JA02	SC182Q10JA01
SC018Q09JA01	SC182Q10JA02
SC018Q09JA02	SCHSIZE
SC018Q10JA01	SMRATIO
SC018Q10JA02	STRATIO

Country Variable Suppression	
SC168Q01JA SC168Q02JA SC168Q03JA SC168Q04JA	STRATUM TOTAT TOTMATH TOTSTAFF
<b>Cyprus</b>	
LANGTEST_COG LANGTEST_QQQ LANGTEST_QQQ SC001Q01TA STRATUM	
<b>Germany</b>	
STRATUM	
<b>Iceland</b>	
GRADE SC002Q01TA SC002Q02TA SC004Q01TA SC013Q01TA SC014Q01TA ST001D01T	ST003D02T ST019AQ01T ST019BQ01T ST019CQ01T ST022Q01TA ST230Q01JA TOTAT
<b>Israel</b>	
STRATUM	
<b>Italy</b>	
REGION STRATUM	
<b>Japan</b>	
IMMIG	
<b>Jordan</b>	
STRATUM	
<b>Macao</b>	
LANGTEST_COG LANGTEST_PAQ LANGTEST_QQQ PRIVATESCH PROGN SC013Q01TA SCHLTYPE	
<b>Malaysia</b>	
SC012Q03TA SC012Q05TA SC012Q08JA SC012Q10JA SC012Q12JA ST038Q09JA ST038Q10JA ST038Q11JA ST261Q02JA	ST261Q03JA ST261Q09JA ST265Q03JA ST265Q04JA ST266Q02JA ST266Q03JA ST266Q04JA ST266Q05JA
<b>Montenegro</b>	
SC013Q01TA ST003D02T	
<b>New Zealand</b>	
SC002Q01TA SC002Q02TA SC004Q01TA SC004Q02TA SC018Q01TA01 SC018Q01TA02	SC182Q06WA01 SC182Q06WA02 SC182Q07JA01 SC182Q07JA02 SC182Q08JA01 SC182Q08JA02

Country Variable Suppression	
SC018Q02TA01	SC182Q09JA01
SC018Q02TA02	SC182Q09JA02
SC018Q08JA01	SC182Q10JA01
SC018Q08JA02	SC182Q10JA02
SC018Q09JA01	SCHSIZE
SC018Q09JA02	TOTAT
SC018Q10JA01	TOTMATH
SC018Q10JA02	TOTSTAFF
SC182Q01WA01	WB151Q01HA
SC182Q01WA02	WB152Q01HA
<b>Norway</b>	
CLSIZE	SC018Q08JA01
GRADE	SC018Q08JA02
LANGTEST_COG	SC018Q09JA01
LANGTEST_QQQ	SC018Q09JA02
MCLSIZE	SC018Q10JA01
PRIVATESCH	SC018Q10JA02
PROADMIN	SC168Q01JA
PROATCE	SC168Q02JA
PROMGMT	SC168Q03JA
PROSTAF	SC168Q04JA
PROPAT6	SC182Q01WA01
PROPAT7	SC182Q01WA02
PROPAT8	SC182Q06WA01
PROPMATH	SC182Q06WA02
PROPSUPP	SC182Q07JA01
RATCMP1	SC182Q07JA02
RATCMP2	SC182Q08JA01
RATTAB	SC182Q08JA02
SC002Q01TA	SC182Q09JA01
SC002Q02TA	SC182Q09JA02
SC004Q01TA	SC182Q10JA01
SC004Q02TA	SC182Q10JA02
SC004Q03TA	SCHLTYPE
SC004Q05NA	SCHSIZE
SC004Q06NA	SMRATIO
SC004Q07NA	ST001D01T
SC004Q08JA	ST003D02T
SC012Q03TA	ST003D03T
SC013Q01TA	STRATIO
SC014Q01TA	TOTAT
SC016Q01TA	TOTMATH
SC016Q02TA	TOTSTAFF
SC016Q03TA	
SC016Q04TA	
SC018Q01TA01	
SC018Q01TA02	
SC018Q02TA01	
SC018Q02TA02	
<b>Singapore</b>	
LANGN	
OCOD1 (2-digit)	
OCOD2 (2-digit)	
SC211Q02JA	
SC211Q03JA	
<b>Sweden</b>	
GRADE	SC182Q06WA02
SC001Q01TA	SC182Q07JA01

Country Variable Suppression	
SC002Q01TA	SC182Q07JA02
SC002Q02TA	SC182Q08JA01
SC004Q01TA	SC182Q08JA02
SC004Q02TA	SC182Q09JA01
SC004Q03TA	SC182Q09JA02
SC004Q08JA	SC182Q10JA01
SC013Q01TA	SC182Q10JA02
SC014Q01TA	SC211Q01JA
SC018Q01TA01	SC211Q02JA
SC018Q01TA02	SC211Q03JA
SC018Q02TA01	SC211Q04JA
SC018Q02TA02	SC211Q05JA
SC018Q08JA01	SC211Q06JA
SC018Q08JA02	ST001D01T
SC018Q09JA01	ST003D02T
SC018Q09JA02	ST003D03T
SC018Q10JA01	ST021Q01TA
SC018Q10JA02	ST022Q01TA
SC182Q01WA01	ST126Q01TA
SC182Q01WA02	ST226Q01JA
SC182Q06WA01	
<b>Thailand</b>	
STRATUM	

Note: 1. Cyprus data are suppressed from the public use files. Information on data for Cyprus: <https://oe.cd/cyprus-disclaimer>.

# 13 Sampling Outcomes

This chapter reports on the PISA 2022 sampling outcomes. Details of the sample design and selection are provided in Chapter 6 of this Technical Report.

## Population coverage

Quality indicators for population coverage and the information used to develop them are presented in Annex Table 13.A.1 and Annex Table 13.A.2, for participating countries/economies and adjudicated regions, respectively. 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.

The components used for the coverage indices are the following:

- ST7a\_1: National population of all 15-year-olds based on national statistics.
- ST7a\_2.1: Enrolled 15-year-old students in grades 7 and above based on national statistics.
- ST7b\_1: Target population that includes all enrolled 15-year-old students in grades 7 and above that omits schools based on national statistics such as schools located in unsafe areas.
- ST7b\_3: Target population that includes all enrolled 15-year-old students in grades 7 and above, minus school-level exclusions, based on national statistics.
- P: Weighted number of participating students calculated from the PISA sample.
- E: Weighted estimate of within-school excluded students calculated from the PISA sample.
- S: Estimate of enrolled students from school sampling frame 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.

**Coverage Index 1:** Coverage of the national *target* population,  $P/(P+E) \times (ST7b_3/ST7b_1)$ . This estimates the extent to which the weighted participants covered the final *target* population after all exclusions. It indicates the overall proportion of the *target* population covered by the non-excluded portion of the student sample.

**Coverage Index 2:** Coverage of the national *enrolled* population,  $P/(P+E) \times (ST7b_3/ST7a_2.1)$ . This estimates the extent to which the weighted participants covered the population of all *enrolled* students in grades 7 and above. Thus, this index may be somewhat lower than Index 1.

**Coverage Index 3:** Coverage of the national *15-year-old* population,  $P/ST7a_1$ . This estimates the proportion of the national population of 15-year-olds covered by the non-excluded portion of the student sample. It is below 1.0 to the extent that 15-year-olds were excluded, or not enrolled in grade 7 or higher.

**Coverage Index 4:** Coverage of the estimated school population,  $(P+E)/S$ . This estimates the proportion of the estimated school 15-year-old population that is represented by the weighted student sample of all PISA-eligible 15-year-old students. Its purpose is to assess whether the enrolment data on the sampling frame is a reliable measure of the number of enrolled 15-year-olds. As the enrolment data on the frame was often inaccurate, this index usually differed noticeably from 1.0. In such cases, Indexes 1 and 2 may be suspect, as they rely on national enrolment data for their denominators, often derived from the same source as the school-level enrolment data.

**Coverage Index 5:** Coverage of the school sampling frame population,  $S/ST7b_3$ . This estimate provides a check as to whether the data on enrolment obtained from national statistics is consistent with the enrolment on the sampling frame. However, in most cases for PISA, the enrolment data based on national statistics were derived using data from the sampling frame by the NPM, and so this ratio was close to 1.0 for most countries/economies, even when the enrolment data on the school sampling frame were poor.

Annex Table 13.A.3, Annex Table 13.A.4, Annex Table 13.A.5, Annex Table 13.A.6, Annex Table 13.A.7 and Annex Table 13.A.8 present school and student-level response rates at the national and regional levels. Response rates are all presented separately by participating country/economy, and by adjudicated regions.

When 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 33%). 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 33% 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.

When 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 33%). 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 33% 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.

When 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/economies 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:

$$\text{weighted school response rate before replacement} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i} \quad \text{Equation 13.1}$$

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. The weighted school response rate, after replacement, is given by the formula:

$$\text{weighted school response rate after replacement} = \frac{\sum_{i \in (Y \cup R)} W_i E_i}{\sum_{i \in (Y \cup R \cup N)} W_i E_i} \quad \text{Equation 13.2}$$

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. 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 students appear in the numerator and denominator as for unweighted rates, but each student is 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/economies with no oversampling of any explicit strata, weighted and unweighted student response 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.

## Teacher response rates

Unweighted response rates for 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.

For weighted teacher response rates, the same teachers appear in the numerator and denominator as for unweighted rates, but each teacher is weighted by its teacher base weight. This is given as the product of the school base weight – for the school in which the teacher was working – and the reciprocal of the teacher selection probability within the school (Annex Table 13.A.).

## 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/economy, 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.

Therefore, a simple random sample of 4 000 students within a country/economy 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:

$$\sigma_{(\hat{\mu})} = \sqrt{\frac{\sigma^2}{n}} \quad \text{Equation 13.3}$$

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:

$$\sigma_{(\hat{\mu})} = \sqrt{\frac{\sigma_{schools}^2}{n_{schools}} + \frac{\sigma_{within}^2}{n_{schools}n_{students}}} \quad \text{Equation 13.4}$$

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:

$$Rho = \frac{\sigma_{schools}^2}{\sigma_{schools}^2 + \sigma_{within}^2}$$
Equation 13.5

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.

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<sub>[1]</sub>).

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:

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)}$$
Equation 13.6

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 a hypothetical country/economy, where the unbiased *BRR* standard error on the mean proficiency estimate is equal to 1.46, and the standard deviation is equal to 102.29, on a sample of 14 530 students, the design effect for the mean proficiency estimate is therefore calculated as:

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} = \frac{(1.46)^2}{[102.29^2/14\ 530]} = 2.96$$
Equation 13.7

This means the sampling variance on the proficiency estimate is about 2.96 times larger than it would have been with a simple random sample of the same sample size.

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:

Equation 13.8

$$Effn(t) = \frac{n}{Deff(t)} = \frac{n \leftarrow Var_{SRS}(t)}{Var_{BRR}(t)}$$

where  $n$  is equal to the actual number of units in the sample. The effective sample size in our example would then be equal to:

$$Effn(t) = \frac{n}{Deff(t)} = \frac{14\,530}{2.96} = 4\,909$$

Equation 13.9

In other words, a simple random sample of about 4 909 students in this hypothetical country/economy would have been as precise as the actual sample for the national proficiency estimate.

### 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/economies, 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 can be 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 population estimates based on performance variables (scale scores) in the international PISA reports consist of two components: sampling variance ( $Var_{BRR}$ ) and measurement variance. The measurement variance is approximated by means of the imputation variance ( $MVar$ ) which is calculated from the statistics calculated from imputed plausible values assigned to the participating students.

The standard error of proficiency estimates in PISA are inflated because the students were not sampled according to a simple random sample and because the estimation of student proficiency includes some amount of measurement error.

Therefore, the variance of a statistic calculated using plausible values is then calculated as the sum of the sampling and the imputation variances, or  $Var_{BRR} + MVar$ .

The five design effects and their respective effective sample sizes can then be defined as follows:

**Design Effect 1:** 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 simple random samples.

$$Def f_1(r) = \frac{Var_{SRS}(r) + MVar(r)}{Var_{SRS}(r)} \quad \text{Equation 13.10}$$

**Design Effect 2:** shows the inflation of the *total* variance due only to the use of a complex sampling design.

$$Def f_2(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r) + MVar(r)} \quad \text{Equation 13.11}$$

**Design Effect 3:** shows the inflation of the sampling variance due to the use of a complex design. This is the same as Formula 7 introduced above.

$$Def f_3(r) = \frac{Var_{BRR}(r)}{Var_{SRS}(r)} \quad \text{Equation 13.12}$$

**Design Effect 4:** shows the inflation of the total variance due to measurement variance.

$$Def f_4(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{BRR}(r)} \quad \text{Equation 13.13}$$

**Design Effect 5:** shows the inflation of the total variance due to the measurement variance and due to the complex sampling design.

$$Def f_5(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r)} \quad \text{Equation 13.14}$$

Table 13.10, Table 13.11., Table 13.12. Table 13.13, Table 13.14. Table 13.15. Table 13.16 present the values of the different design effects and the effective sample size using  $Def f_5$ , for each of the main PISA domains.

To better understand the design effect for a country/economy, some information related to the design effects and their respective effective sample sizes are presented in Annex C.

## References

Cochran, W. (1977), *Sampling Techniques (3rd ed.)*, John Wiley and Sons. [1]

## Annex 13.A. Sampling outcomes

Web tables for each chapter can be accessed via the StatLink.

### Annex Table 13.A.1. Chapter 13: Population Characteristics, Response Rates, and Proficiency Errors Across Regions and Domains

Tables	Title
Web Table 13.A.2	Population characteristics, sample characteristics, exclusions and coverage indices for participating countries/economies
Web Table 13.A.3	Population characteristics, sample characteristics, exclusions and coverage indices for participating adjudicated regions
Table 13.A.4	Response rates for participating countries/economies calculated by using only original schools and no replacement schools
Table 13.A.5	Response rates for adjudicated regions calculated by using only original schools and no replacement schools
Table 13.A.6	Response rates for participating countries/economies when first and second replacement schools were accounted for in the rates
Table 13.A.7	Response rates for adjudicated regions when first and second replacement schools were accounted for in the rates
Table 13.A.8	Student response rates among the full set of participating schools in participating countries/economies
Table 13.A.9	Student response rates among the full set of participating schools in adjudicated regions
Table 13.A.10	Teacher response rates among the full set of participating schools in participating countries
Web Table 13.A.11	Standard errors and related statistics for the average mathematics proficiency
Web Table 13.A.12	Standard errors and related statistics for the average reading proficiency
Web Table 13.A.13	Standard errors and related statistics for the average science proficiency
Web Table 13.A.14	Standard errors and related statistics for the average creative thinking proficiency
Web Table 13.A.15	Standard errors and related statistics for the average financial literacy proficiency (Financial Literacy sample)
Web Table 13.A.16	Standard errors and related statistics for the average mathematics proficiency (Financial Literacy sample)
Web Table 13.A.17	Standard errors and related statistics for the average reading proficiency (Financial Literacy sample)

### Annex Table 13.A.4. Response rates for participating countries/economies calculated by using only original schools and no replacement schools

Country/economy	Weighted School Participation Rate Before Replacement (%) (SCHRRW1)	Weighted Number of Responding Schools (Weighted also by enrolment) (NUMW1)	Weighted Number of Schools Sampled (responding + nonresponding) (Weighted also by enrolment) (DENW1)	Unweighted School Participation Rate Before Replacement (%) (SCHRRU1)	Number of Responding Schools (Unweighted) (NUMU1)	Number of Responding and Nonresponding Schools (Unweighted) (DENU1)
Albania	94.7	27530	29067	93.2	274	294
Argentina	98.3	661503	673069	98.5	454	461
Australia	92.5	260643	281781	90.9	722	794
Austria	95.7	77289	80733	94.3	300	318
Baku (Azerbaijan)	100.0	31925	31925	100.0	178	178
Belgium	80.3	101303	126138	76.4	243	318
Brazil	80.9	2153176	2660537	79.4	505	636
Brunei	100.0	6675	6675	100.0	54	54
Bulgaria	84.5	47378	56052	85.5	177	207
Cambodia	99.6	205960	206763	99.5	182	183
Canada	81.3	305746	375877	78.9	828	1049

Chile	84.3	187116	222091	82.0	205	250
Chinese Taipei	82.6	161354	195232	83.3	180	216
Colombia	96.6	658016	681141	94.3	249	264
Costa Rica	99.0	64480	65122	99.0	198	200
Croatia	99.8	37398	37475	98.9	180	182
Cyprus	97.5	8875	9100	96.2	101	105
Czechia	100.0	98609	98609	100.0	430	430
Denmark	90.1	53540	59431	87.6	325	371
Dominican Republic	98.5	131827	133900	96.9	249	257
El Salvador	99.6	73847	74135	99.0	288	291
Estonia	99.4	13659	13745	98.5	196	199
Finland	99.5	60180	60501	98.4	241	245
France	99.6	790568	794003	99.6	282	283
Georgia	93.6	40653	43421	93.3	250	268
Germany	92.9	674828	726200	91.3	241	264
Greece	90.1	90812	100785	89.7	217	242
Guatemala	85.0	143290	168547	73.4	265	361
Hong Kong (China)	59.6	32428	54402	59.8	122	204
Hungary	88.8	82009	92393	89.2	249	279
Iceland	96.4	4435	4601	89.9	134	149
Indonesia	99.3	3985101	4011189	99.3	408	411
Ireland	99.4	68814	69234	99.4	169	170
Israel	90.7	124237	137007	89.5	188	210
Italy	96.0	493350	513656	95.4	334	350
Jamaica	89.8	41020	45680	89.0	145	163
Japan	91.9	949447	1033001	91.5	182	199
Jordan	100.0	146365	146365	100.0	260	260
Kazakhstan	98.5	279305	283489	98.9	565	571
Korea	88.9	369002	415104	88.8	166	187
Kosovo	96.1	23183	24127	91.2	229	251
Latvia	83.9	15494	18464	80.3	208	259
Lithuania	99.6	25311	25418	98.3	288	293
Macao (China)	100.0	4453	4453	100.0	46	46
Malaysia	99.7	406803	407861	99.5	199	200
Malta	100.0	4114	4114	100.0	46	46
Mexico	95.9	1473466	1535688	94.1	272	289
Mongolia	100.0	43631	43631	100.0	195	195
Montenegro	98.8	6581	6659	98.4	63	64
Morocco	99.8	479666	480608	99.4	177	178
Netherlands	65.5	116517	177833	65.1	114	175
New Zealand	61.4	35524	57847	61.7	140	227
North Macedonia	100.0	17919	17919	100.0	111	111
Norway	98.7	62129	62943	98.2	266	271
Palestinian Authority	99.0	94105	95053	98.9	271	274
Panama	84.1	54532	64834	78.2	190	243
Paraguay	98.7	87772	88922	97.9	278	284
Peru	94.0	489130	520113	91.1	308	338
Philippines	100.0	1719012	1719012	100.0	188	188
Poland	88.6	309061	348856	88.5	223	252
Portugal	94.7	95312	100641	93.8	213	227
Qatar	100.0	18927	18927	100.0	229	229
Republic of Moldova	99.7	29607	29687	98.9	265	268
Romania	100.0	167589	167589	100.0	262	262

Saudi Arabia	91.9	300026	326333	91.3	178	195
Serbia	98.7	63599	64435	96.8	183	189
Singapore	98.5	41915	42567	98.2	164	167
Slovak Republic	90.5	44081	48692	90.0	271	301
Slovenia	97.2	18729	19264	91.7	344	375
Spain	97.7	473996	485037	97.4	959	985
Sweden	97.8	113994	116574	96.6	259	268
Switzerland	95.1	73464	77247	93.3	249	267
Thailand	98.8	685471	693755	98.6	276	280
Türkiye	99.4	1079992	1086638	99.5	195	196
Ukraine (18 of 27 Regions)	79.8	178606	223859	74.6	141	189
United Arab Emirates	99.8	63395	63507	99.6	840	843
United Kingdom	67.3	490313	728369	66.9	388	580
United States of America	51.4	2019439	3927302	49.4	125	253
Uruguay	99.4	43188	43447	99.1	221	223
Uzbekistan	100.0	510406	510406	100.0	202	202
Viet Nam	100.0	1020528	1020528	100.0	178	178

**Annex Table 13.A.5. Response rates for adjudicated regions calculated by using only original schools and no replacement schools**

Country/economy	Weighted School Participation Rate Before Replacement (%) (SCHRRW1)	Weighted Number of Responding Schools (Weighted also by enrolment) (NUMW1)	Weighted Number of Schools Sampled (responding + nonresponding) (Weighted also by enrolment) (DENW1)	Unweighted School Participation Rate Before Replacement (%) (SCHRRU1)	Number of Responding Schools (Unweighted) (NUMU1)	Number of Responding and Nonresponding Schools (Unweighted) (DENU1)
Argentina (CABA)	100.0	38009	38009	100.0	80	80
Argentina (Córdoba)	98.7	53002	53675	98.8	83	84
Argentina (Mendoza)	100.0	30381	30381	100.0	92	92
Belgium (Flanders)	71.8	51049	71073	67.8	135	199
United Arab Emirates (Abu Dhabi)	100.0	23381	23381	100.0	289	289
United Arab Emirates (Dubai)	99.4	17548	17660	98.8	250	253
United Arab Emirates (Sharjah)	100.0	13232	13232	100.0	183	183
United Kingdom (Scotland)	87.5	51700	59080	87.6	106	121

**Annex Table 13.A.6. Response rates for participating countries/economies when first and second replacement schools were accounted for in the rates**

Country/economy	Weighted School Participation Rate After All Replacements (%) (SCHRRW3)	Weighted Number of Responding Schools (Weighted also by enrolment) (NUMW3)	Weighted Number of Schools Sampled (responding + nonresponding) (Weighted also by enrolment) (DENW3)	Unweighted School Participation Rate After All Replacements (%) (SCHRRU3)	Number of Responding Schools (Unweighted) (NUMU3)	Number of Responding and Nonresponding Schools (Unweighted) (DENU3)
Albania	94.7	27530	29067	93.2	274	294

Argentina	99.2	668001	673236	99.1	457	461
Australia	95.6	269918	282241	93.6	743	794
Austria	96.3	77799	80750	95.0	302	318
Baku (Azerbaijan)	100.0	31925	31925	100.0	178	178
Belgium	91.4	115591	126446	89.6	285	318
Brazil	95.6	2541343	2659664	94.2	599	636
Brunei	100.0	6675	6675	100.0	54	54
Bulgaria	97.7	54795	56079	97.6	202	207
Cambodia	100.0	207046	207046	100.0	183	183
Canada	85.6	321877	376040	82.7	867	1049
Chile	94.2	208702	221439	92.0	230	250
Chinese Taipei	83.8	163590	195232	84.3	182	216
Colombia	99.2	683439	688995	99.2	262	264
Costa Rica	99.0	64480	65122	99.0	198	200
Croatia	99.8	37398	37475	98.9	180	182
Cyprus	97.5	8875	9100	96.2	101	105
Czechia	100.0	98609	98609	100.0	430	430
Denmark	96.2	57254	59517	93.5	347	371
Dominican Republic	99.4	133159	133900	98.4	253	257
El Salvador	99.9	74136	74212	99.7	290	291
Estonia	99.4	13659	13745	98.5	196	199
Finland	99.5	60180	60501	98.4	241	245
France	99.6	790568	794003	99.6	282	283
Georgia	99.8	43539	43611	99.6	267	268
Germany	98.2	712724	725905	97.3	257	264
Greece	96.1	96821	100772	95.0	230	242
Guatemala	92.6	155960	168475	80.3	290	361
Hong Kong (China)	79.9	43491	54402	79.9	163	204
Hungary	98.6	90673	91964	96.8	270	279
Iceland	96.4	4435	4601	89.9	134	149
Indonesia	99.8	4002841	4011189	99.8	410	411
Ireland	100.0	69234	69234	100.0	170	170
Israel	92.9	127287	137007	91.9	193	210
Italy	99.4	510819	513842	98.6	345	350
Jamaica	90.9	41545	45680	90.2	147	163
Japan	91.9	949447	1033001	91.5	182	199
Jordan	100.0	146365	146365	100.0	260	260
Kazakhstan	100.0	283481	283481	100.0	571	571
Korea	99.7	413724	415104	99.5	186	187
Kosovo	96.1	23183	24127	91.2	229	251
Latvia	88.7	16424	18516	86.9	225	259
Lithuania	100.0	25408	25414	99.7	292	293
Macao (China)	100.0	4453	4453	100.0	46	46
Malaysia	99.7	406803	407861	99.5	199	200
Malta	100.0	4114	4114	100.0	46	46
Mexico	98.9	1519261	1535688	96.9	280	289
Mongolia	100.0	43631	43631	100.0	195	195
Montenegro	98.8	6581	6659	98.4	63	64
Morocco	100.0	479939	479939	100.0	178	178
Netherlands	89.6	159228	177613	88.0	154	175
New Zealand	72.4	41871	57865	74.4	169	227
North Macedonia	100.0	17919	17919	100.0	111	111
Norway	99.1	62393	62943	98.5	267	271



Palestinian Authority	100.0	94988	95027	99.6	273	274
Panama	91.3	59341	64996	88.5	215	243
Paraguay	99.6	88602	88922	98.9	281	284
Peru	99.9	521500	522136	99.7	337	338
Philippines	100.0	1719012	1719012	100.0	188	188
Poland	96.1	335389	348856	95.2	240	252
Portugal	99.2	99768	100578	98.7	224	227
Qatar	100.0	18927	18927	100.0	229	229
Republic of Moldova	99.7	29607	29687	98.9	265	268
Romania	100.0	167589	167589	100.0	262	262
Saudi Arabia	99.6	325174	326372	99.0	193	195
Serbia	98.7	63599	64435	96.8	183	189
Singapore	98.5	41915	42567	98.2	164	167
Slovak Republic	95.5	46387	48549	95.7	288	301
Slovenia	97.3	18747	19264	92.0	345	375
Spain	99.1	480541	485037	98.1	966	985
Sweden	98.9	115248	116574	97.8	262	268
Switzerland	98.2	76060	77488	97.0	259	267
Thailand	99.5	690286	693755	99.6	279	280
Türkiye	100.0	1086638	1086638	100.0	196	196
Ukraine (18 of 27 Regions)	91.0	204043	224119	86.8	164	189
United Arab Emirates	99.8	63395	63507	99.6	840	843
United Kingdom	81.8	593600	725986	77.8	451	580
United States of America	63.3	2485876	3926991	60.9	154	253
Uruguay	99.9	43395	43447	99.6	222	223
Uzbekistan	100.0	510406	510406	100.0	202	202
Viet Nam	100.0	1020528	1020528	100.0	178	178

**Annex Table 13.A.7. Response rates for adjudicated regions when first and second replacement schools were accounted for in the rates**

Country/economy	Weighted School Participation Rate After All Replacements (%) (SCHRRW3)	Weighted Number of Responding Schools (Weighted also by enrolment) (NUMW3)	Weighted Number of Schools Sampled (responding + nonresponding) (Weighted also by enrolment) (DENW3)	Unweighted School Participation Rate After All Replacements (%) (SCHRRU3)	Number of Responding Schools (Unweighted) (NUMU3)	Number of Responding and Nonresponding Schools (Unweighted) (DENU3)
Argentina (CABA)	100.0	38009	38009	100.0	80	80
Argentina (Córdoba)	98.7	53002	53675	98.8	83	84
Argentina (Mendoza)	100.0	30381	30381	100.0	92	92
Belgium (Flanders)	88.6	63321	71477	86.4	172	199
United Arab Emirates (Abu Dhabi)	100.0	23381	23381	100.0	289	289
United Arab Emirates (Dubai)	99.4	17548	17660	98.8	250	253
United Arab Emirates (Sharjah)	100.0	13232	13232	100.0	183	183
United Kingdom (Scotland)	96.4	57164	59316	96.7	117	121

**Annex Table 13.A.8. Student response rates among the full set of participating schools in participating countries/economies**

Country/economy	Weighted Student Participation Rate After All Replacements (%) (STURRW3)	Number of Students Assessed (Weighted) (NUMSTW3)	Number of Students Sampled (assessed + absent) (Weighted) (DENSTW3)	Unweighted Student Participation Rate After All Replacements (%) (STURRU3)	Number of Students Assessed (Unweighted) (NUMSTU3)	Number of Students Sampled (assessed + absent) (Unweighted) (DENSTU3)
Albania	86.5	23274	26915	86.5	6129	7089
Argentina	85.8	508035	592257	86.4	12111	14014
Australia	76.1	193102	253899	75.6	13437	17771
Austria	88.8	65057	73230	86.7	6151	7092
Baku (Azerbaijan)	87.8	26799	30529	87.8	7720	8793
Belgium	86.6	101344	117082	86.9	8286	9533
Brazil	84.2	1832626	2177600	83.8	10798	12879
Brunei	93.2	5576	5980	93.2	5576	5980
Bulgaria	88.8	46335	52192	88.79	6107	6878
Cambodia	99.4	125643	126409	99.45	5279	5308
Canada	77.0	233773	303622	78.93	23073	29234
Chile	84.0	168773	201037	85.07	6488	7627
Chinese Taipei	82.3	131517	159821	83.22	5857	7038
Colombia	91.8	532284	580114	92.15	7804	8469
Costa Rica	92.0	52220	56750	91.84	6113	6656
Croatia	85.2	29804	34963	85.28	6135	7194
Cyprus	83.8	7190	8578	83.90	6515	7765
Czechia	91.2	91518	100330	91.14	8460	9282
Denmark	84.2	46126	54775	83.17	6200	7455
Dominican Republic	92.7	112417	121281	92.60	6868	7417
El Salvador	93.6	63767	68101	93.67	6705	7158
Estonia	88.2	11693	13262	88.34	6392	7236
Finland	88.7	52007	58641	86.69	10239	11811
France	90.7	705197	777730	90.16	6770	7509
Georgia	98.1	39587	40348	98.08	6583	6712
Germany	88.0	588741	669277	87.82	6116	6964
Greece	92.4	87038	94215	92.52	6403	6921
Guatemala	91.4	143084	156600	90.91	5190	5709
Hong Kong (China)	75.3	29278	38858	75.55	5907	7819
Hungary	92.3	80160	86877	92.44	6198	6705
Iceland	80.1	3360	4195	80.10	3360	4195
Indonesia	95.2	3602554	3782864	95.72	13439	14040
Ireland	76.8	50274	65497	76.73	5569	7258
Israel	84.1	103556	123165	84.05	6251	7437
Italy	91.9	452653	492440	92.33	10552	11429
Jamaica	67.6	15622	23123	66.88	3873	5791
Japan	91.9	858514	934656	91.57	5760	6290
Jordan	97.5	140640	144269	97.32	7799	8014
Kazakhstan	98.3	267773	272446	98.22	19769	20128
Korea	94.4	383999	406986	94.4	6454	6840
Kosovo	91.1	18427	20220	91.1	6027	6616
Latvia	88.5	13215	14935	88.6	5373	6067
Lithuania	92.7	22470	24245	92.7	7257	7826
Macao (China)	99.1	4384	4423	99.1	4384	4423

Malaysia	93.5	362809	387928	93.6	7069	7554
Malta	79.1	3127	3955	79.1	3127	3955
Mexico	94.9	1313477	1383827	94.2	6288	6675
Mongolia	97.9	39969	40828	97.8	6999	7155
Montenegro	94.6	5954	6291	94.7	5793	6117
Morocco	98.1	446431	454986	98.1	6867	7000
Netherlands	80.9	113351	140125	81.1	5046	6221
New Zealand	71.7	29219	40758	71.3	4682	6567
North Macedonia	89.6	14832	16548	89.6	6610	7380
Norway	86.7	50577	58362	86.6	6611	7635
Palestinian Authority	96.2	85017	88348	95.9	7905	8239
Panama	76.8	29491	38418	75.5	4544	6017
Paraguay	92.0	74217	80700	92.1	5084	5522
Peru	97.5	486292	498888	97.6	6968	7136
Philippines	95.2	1698135	1782896	95.3	7193	7550
Poland	81.0	266114	328452	81.0	6011	7422
Portugal	86.1	82496	95838	86.1	6793	7888
Qatar	89.0	16346	18361	88.8	7676	8649
Republic of Moldova	94.1	27114	28799	94.1	6235	6623
Romania	97.4	157838	162019	97.6	7364	7543
Saudi Arabia	97.1	307363	316501	97.0	6928	7144
Serbia	91.2	53150	58297	91.2	6413	7033
Singapore	91.4	37797	41358	91.3	6606	7235
Slovak Republic	90.9	41319	45438	91.4	5824	6375
Slovenia	82.5	15142	18355	82.6	6721	8134
Spain	86.3	392413	454692	86.8	30800	35472
Sweden	85.1	91230	107261	85.1	6072	7133
Switzerland	90.9	67555	74335	91.4	6829	7471
Thailand	96.4	580014	601524	96.4	8495	8816
Türkiye	98.0	914714	933402	98.1	7250	7387
Ukraine (18 of 27 Regions)	86.9	131271	151104	86.0	3876	4508
United Arab Emirates	92.9	56369	60658	92.5	24600	26592
United Kingdom	75.2	448396	596519	76.2	12972	17023
United States of America	79.9	1866014	2336430	79.6	4552	5719
Uruguay	86.7	35308	40728	86.7	6618	7637
Uzbekistan	98.1	472726	482059	98.0	7293	7445
Viet Nam	99.4	933854	939459	99.4	6068	6105

**Annex Table 13.A.9. Student response rates among the full set of participating schools in adjudicated regions**

Country/economy	Weighted Student Participation Rate After All Replacements (%) (STURRW3)	Number of Students Assessed (Weighted) (NUMSTW3)	Number of Students Sampled (assessed + absent) (Weighted) (DENSTW3)	Unweighted Student Participation Rate After All Replacements (%) (STURRU3)	Number of Students Assessed (Unweighted) (NUMSTU3)	Number of Students Sampled (assessed + absent) (Unweighted) (DENSTU3)
Argentina (CABA)	85.5	29481	34493	85.5	2251	2634
Argentina (Córdoba)	89.9	42309	47068	90.5	2217	2449
Argentina (Mendoza)	85.7	26446	30842	86.3	2514	2914
Belgium (Flanders)	87.4	55935	63968	87.4	4714	5393
United Arab Emirates (Abu Dhabi)	92.6	20493	22132	92.2	8316	9017

United Arab Emirates (Dubai)	91.4	15404	16856	90.9	7374	8113
United Arab Emirates (Sharjah)	94.7	11122	11747	94.2	5239	5560
United Kingdom (Scotland)	79.4	39590	49889	79.1	3257	4115

**Annex Table 13.A.10. Teacher response rates among the full set of participating schools in participating countries**

Country/economy	Weighted Teacher Participation Rate After All Replacements (%) (TCHRRW3)	Number of Teachers Assessed (Weighted) (NUMTQW3)	Number of Teachers Sampled (assessed + absent) (Weighted) (DENTQW3)	Unweighted Teacher Participation Rate After All Replacements (%) (TCHRRU3)	Number of Teachers Assessed (Unweighted) (NUMTCH3)	Number of Teachers Sampled (assessed + absent) (Unweighted) (DENTCH3)
Australia	80.6	56269.97	69781.97	80.1	11397	14223
Baku (Azerbaijan)	48.6	7409.80	15249.37	51.3	1915	3736
Brazil	75.5	295437.52	391530.17	77.2	5646	7310
Colombia	87.4	112921.28	129161.43	88.5	2615	2956
Costa Rica	86.6	12440.54	14369.05	87.2	2476	2841
Dominican Republic	58.6	25934.14	44254.82	62.7	2179	3473
Georgia	89.1	18848.29	21154.93	88.1	3202	3635
Germany	72.0	155734.55	216273.34	73.5	3631	4940
Hong Kong (China)	74.1	9148.35	12341.55	75.0	2335	3113
Korea	92.1	176093.25	191096.45	92.5	3614	3906
Kosovo	74.4	2286.30	3072.22	74.1	1290	1741
Macao (China)	99.6	1916.00	1923.00	99.6	1916	1923
Malaysia	99.8	87458.39	87598.58	99.8	3956	3964
Morocco	95.4	93849.95	98337.49	95.7	2998	3134
Panama	78.5	5639.86	7184.30	82.4	1597	1937
Peru	99.2	138807.45	139990.72	99.1	3708	3740
Portugal	92.1	26253.77	28515.19	92.6	3487	3767
United Arab Emirates	85.6	13420.60	15675.31	85.4	10092	11819
Abu Dhabi (United Arab Emirates)	89.5	4932.80	5510.47	89.6	3832	4278
Dubai (United Arab Emirates)	82.3	4118.53	5003.07	81.9	2809	3429
Sharjah (United Arab Emirates)	81.6	2271.97	2784.27	81.4	1961	2410

# 14

## Scaling outcomes

This chapter reports on the outcomes of implementing the item response theory (IRT) and population modelling methods described in Chapter 11 for the PISA 2022 main survey cognitive assessment data. It provides results of the assessments of the invariance of the IRT item parameters within and across countries/economies, estimates of the reliability and correlations across assessments domains, and estimates of the linking errors between the 2022 and prior PISA cycles. The location of the items across the full range of proficiencies based on their common international parameters are also reported. Finally, the correlations between scales and the percentage of students in each country at each proficiency level are presented for each cognitive domain.

### IRT scaling outcomes

IRT scaling outcomes include the proportions of item were invariant across countries and PISA cycles, as well as the common and unique items parameters and the dropped items used for the population modelling of each country/economy data. The international (common) item parameters are provided in this technical report's Annex A and unique country/economy's item parameters are provided in Annex F. The next section provides an assessment of item parameter invariance across countries/economies supporting that the comparability of the PISA scales across cycles and countries was achieved in each domain by reaching a desirable proportion of invariant item parameters across countries/economies and cycles. The following section describes the international characteristics of each domain's item pool and shows the item maps that locate the items on the reporting scales.

### *Invariance of item parameters*

The item parameters for all the items used in the assessment were obtained through IRT scaling. In PISA 2022, IRT scaling was implemented through a multi-group (i.e., country-by-language groups) IRT concurrent calibration using the 2022 main survey data, using the trend items as fixed linking items and setting the scale to the PISA scale established in 2015 and 2018. That is, item parameters for trend items were fixed to the ones used in PISA 2018 (either common international or unique to a specific country-by-language group or groups), unless there was evidence that the 2018 parameters did not fit the 2022 data (see Chapter 11 for details).

In most cases the international item parameters fitted data for all country-by-language groups. When they did not fit a particular country-by-language group, unique or group-specific parameters were estimated and used, unless it was found that the unique parameters could not be estimated, still did not fit the data well enough, or were extreme. That is, an item was dropped if in the end, its RMSD fit could not be reduced to 0.15 or below, its slope parameter was below 0.1 or its difficulty parameter was larger than 5.0 in absolute value. These criteria were not applied to reading fluency items because they typically are very easy items. In rare cases, items were also be dropped when, despite being checked in the field trial, content and/or translation issues were nonetheless found in the main survey—given feedback from countries/economies, content and psychometric reviews. In even rarer cases, items were dropped entirely (in all countries/economies) if analyses indicated that it did not fit the data collected in the majority of the

countries/economies. In this PISA 2022 cycle: one mathematics, one reading, one financial literacy and six creative thinking items were dropped from all groups <sup>1</sup>.

To assess the invariance of item parameters across country-by-language groups and cycles, items were categorized as:

- *invariant* when common international parameters could be used;
- *group-specific invariant* when the same unique parameters could be used across cycles (applies only to trend items);
- *variant* for all other cases where unique item parameters were estimated (new items) or when unique parameters were estimated that are different from the 2018 parameters (trend items); and
- *dropped* when the item could not be fitted to the data and was dropped for one or more country-by-language groups.

For countries with multiple language groups, the number of invariant, variant, or dropped items were averaged across the different language groups within the country to calculate the proportion of unique item parameters used. Sample weights were used for this calculation. Annex Table 14.A.2 shows the proportions of items categorized as invariant, variant, and dropped, averaged across countries participating in the 2022 computer-based assessment (CBA). The proportion of invariant items was large for all domains, ranging from 76.4% for the reading MSAT items to 93.7% for the reading fluency items. A large proportion of invariant items is critical for ensuring the comparability of scores across countries and cycles. Group-specific invariant items also contribute to the comparability of scores across cycles. The proportion of invariant total (invariant and group-specific invariant) was above 98.5% for all domains but creative thinking at 77.4%. Regarding the dropped category, the proportions were small for all domains (less than 2%).

Annex Table 14.A.3 shows the proportion of items categorized as invariant, variant, and dropped, averaged across countries participating in the new 2022 paper-based assessment (new PBA). The results across the three new PBA participating countries showed somewhat lower proportions of invariance than with CBA. Nevertheless, proportions of total invariant items were above 80% for all domains and few items were dropped for any country.

An overview of the frequencies of invariant, variant, and dropped items for each domain is presented in Figure 14.1, Figure 14.2, Figure 14.3, Figure 14.4 and Figure 14.5 for CBA, new PBA and PBA participating countries/economies. Each country is represented by stacked vertical bars: above the horizontal line at zero, dark green represents the number of items classified as invariant and light green represents the number of group-specific invariant items (only trend items); below the 0 horizontal line, yellow represents the number of variant items<sup>2</sup> and red represents the number of items dropped from scaling. The frequencies of variant and dropped items are shown using negative values to highlight differences between the number of items that contribute to ensuring the comparability of the PISA scales (invariant) and the number of items that do not (variant). The countries are sorted from left to right by increasing number of invariant items, first CBA, new PBA, and PBA countries.

These plots show that while there is some variability across countries, the numbers of invariant item parameters and group-specific invariant item parameters are large enough to ensure the comparability of the proficiency estimates across countries/economies and across cycles.

Figure 14.1. Frequency of invariant, variant, and dropped items for mathematics, by country/economy

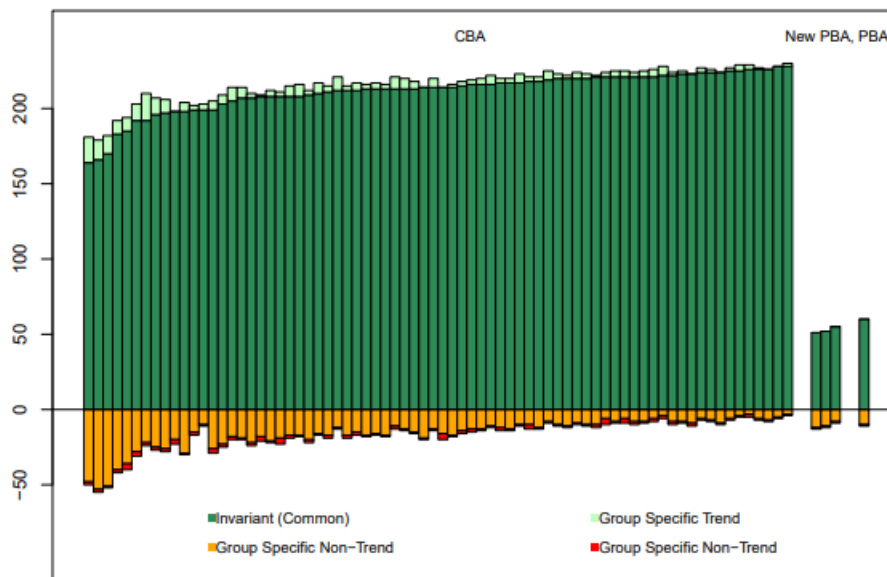
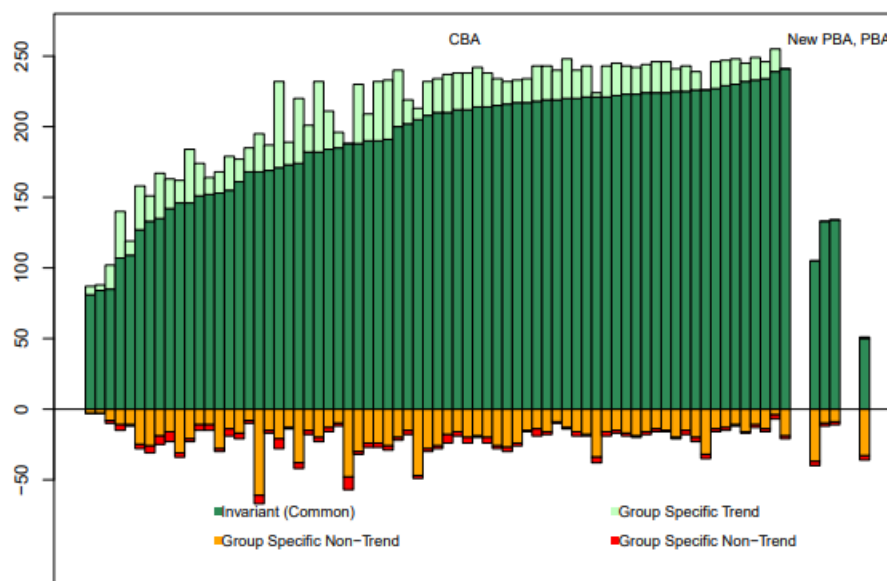


Figure 14.2. Frequency of invariant, variant, and dropped items for reading, by country/economy



Note: Because reading is a minor domain in 2022, in some countries, sample size was not enough to assess fit with the 2022 data. These cases are not included in this plot, resulting in fewer than the number of items used being displayed in these cases. However all items were evaluated for fit in 2018 when reading was the major domain--see PISA 2018 Technical report, Chapter 12 for these results).

Figure 14.3. Frequency of invariant, variant, and dropped items for science, by country/economy

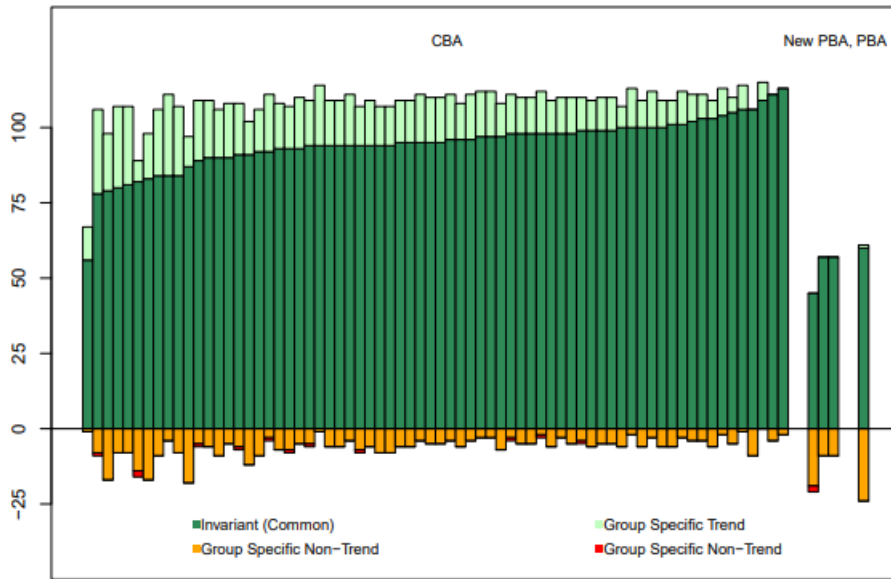
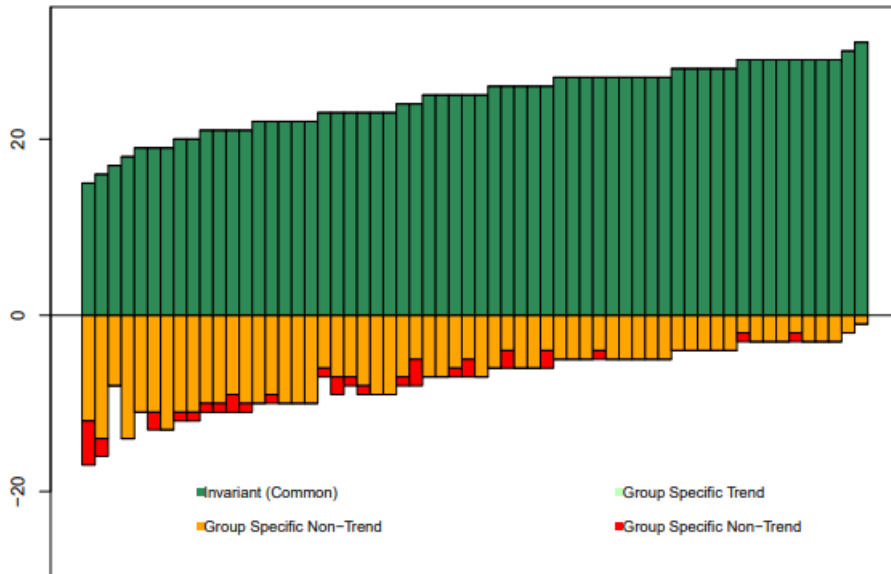
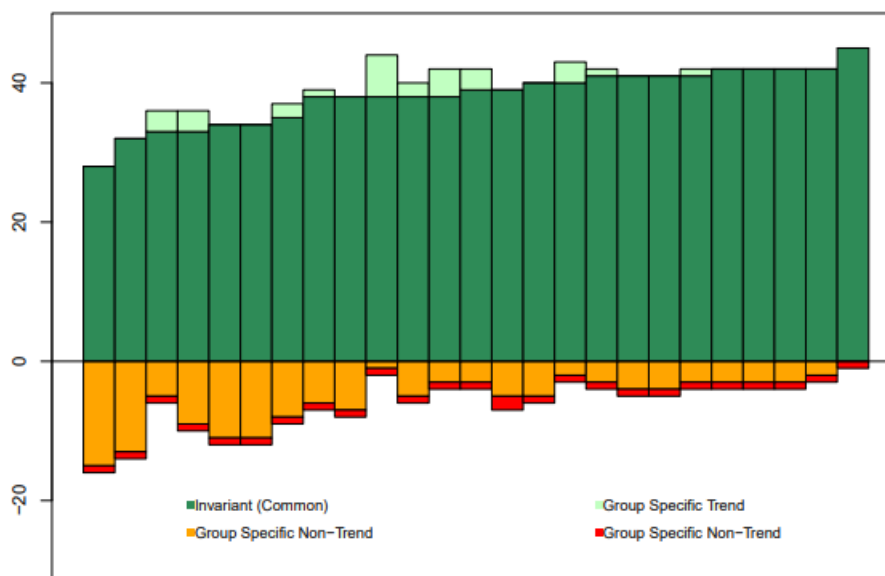


Figure 14.4. Frequency of invariant, variant, and dropped items for creative thinking, by country/economy





**Figure 14.5. Frequency of invariant, variant, and dropped items for financial literacy, by country/economy**



### ***International characteristics of the item pools***

This section provides an overview of the test targeting, the domain inter-correlations, and the correlations among the mathematics subscales.

#### ***Test targeting***

Similar to assigning a specific score on a scale to students according to their performance on an assessment (OECD, 2022<sup>[1]</sup>), each item in PISA 2022 was assigned a specific value on a scale based on response probability (RP) calculated using the item's IRT parameters (discrimination and difficulty). Chapter 17 describes how items can be placed along a scale based on their RP values and how these values can be used to classify items into proficiency levels.

Historically in PISA, a response probability of 0.62 (RP62) has been used to classify items into levels. Students with a proficiency located at or below this point have a probability of 0.62 or less of getting the item correct, while students with a proficiency above this point have a higher probability of getting the item correct higher than 0.62. The RP62 values for all items and their performance level classification are presented in Annex A, together with the final international/common item parameter estimates obtained from the IRT scaling. Note that for polytomous items, the RP62 value is provided for partial credit as well as full credit responses. The partial credit RP62 has been defined as the minimum proficiency level a student need to have an expected score that is 62% of the full credit.

Annex Table 14.A.4, Annex Table 14.A.5, Annex Table 14.A.6, Annex Table 14.A.7 and Annex Table 14.A.8 show the proficiency levels defined for each cognitive domain, along with the percentage of items and the percentages of students classified at each level of proficiency, using the first plausible value. Note that although polytomous items have two RP62 levels (partial credit and full credit), they were classified according to the full credit RP62 only for all domains but creative thinking. For creative thinking, most of the items are polytomous items (28 out of 32), therefore we describe both partial- and full-credit RP62 levels.

Since RP62 values and the plausible values are on the same PISA scale, the distribution of students' latent ability and the items' RP62 values can be compared and contrasted. In Figure 14.6, Figure 14.7, Figure 14.8, Figure 14.9 and Figure 14.10, the left side of each figure illustrates the distribution of the first plausible values (PV1) across countries. In each figure, the blue line indicates the empirical density of the first plausible values across all countries, and the red line indicates the theoretical normal distribution with the mean and the variance of plausible values across all countries. The figures show that the distribution of the plausible values for each domain are approximately normal. On the right side of each figure, the RP62 value for each of the items is plotted. As with the tables above, in all domains but creative thinking, only the RP62 values for full-credit are shown.

**Figure 14.6. Distribution of the first plausible values and item RP62 values in mathematics**

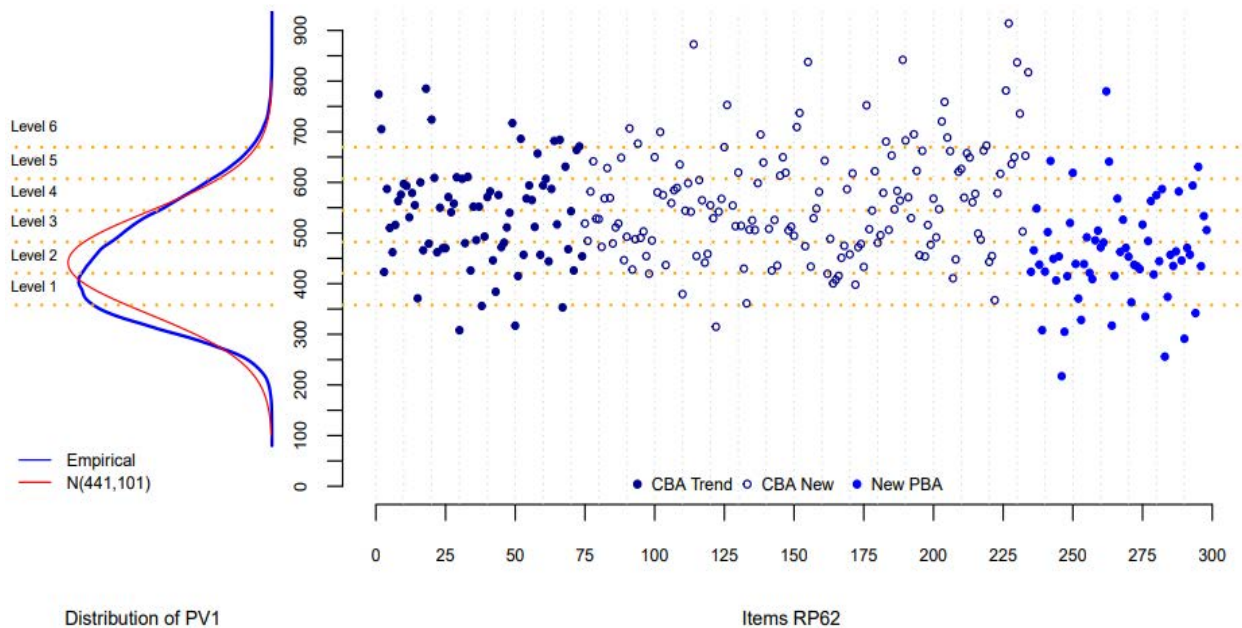


Figure 14.7. Distribution of the first plausible values and item RP62 values in reading

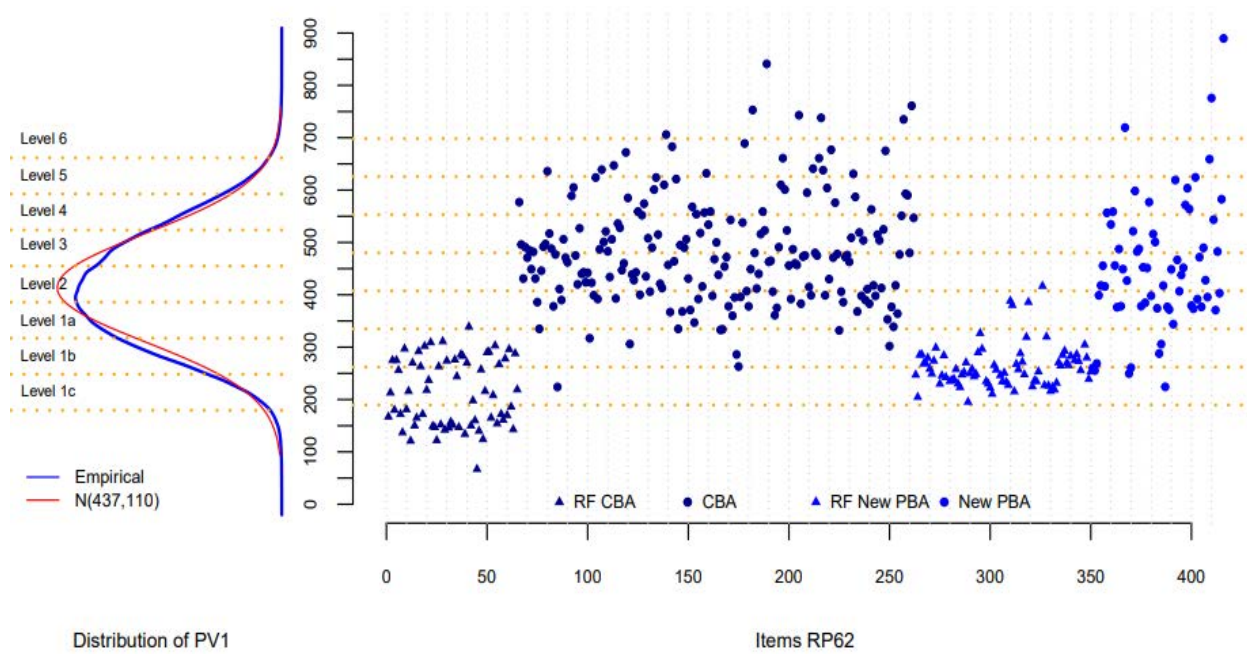


Figure 14.8. Distribution of the first plausible values and item RP62 values in science

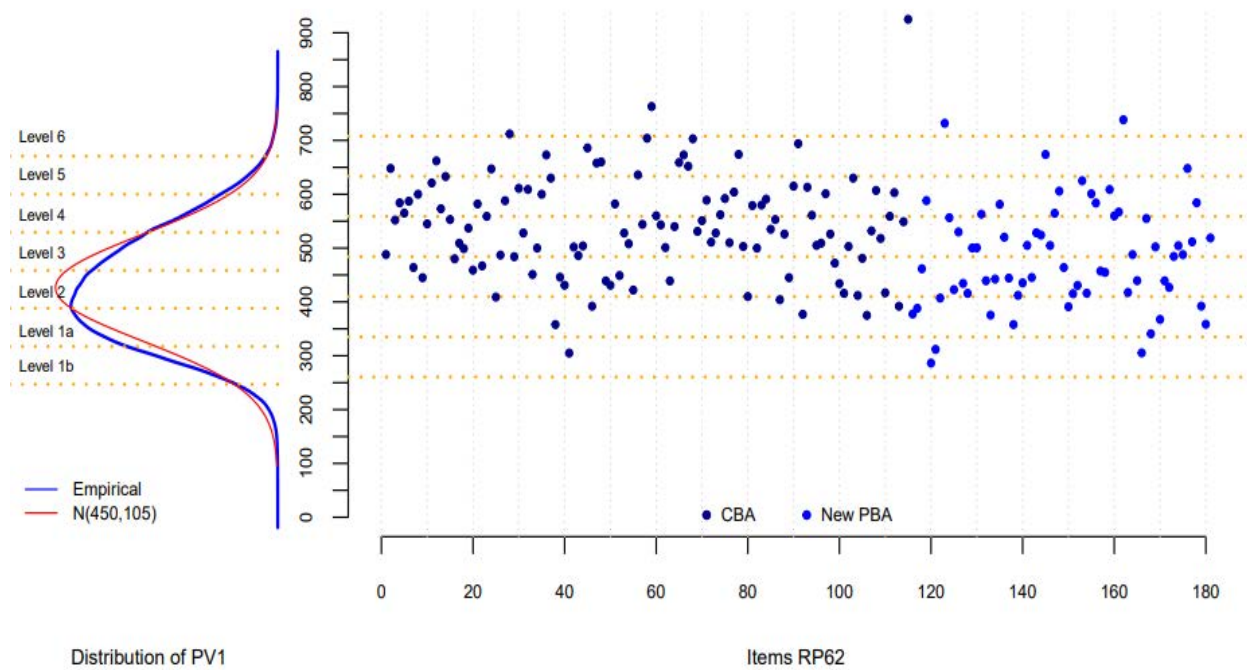


Figure 14.9. Distribution of the first plausible values and item RP62 values in creative thinking

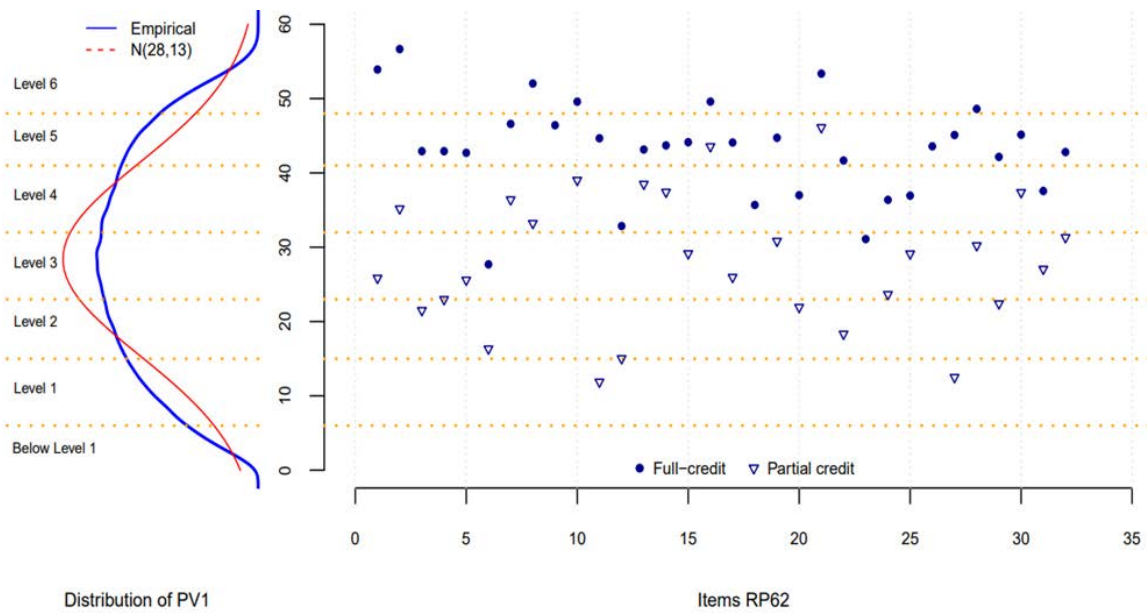
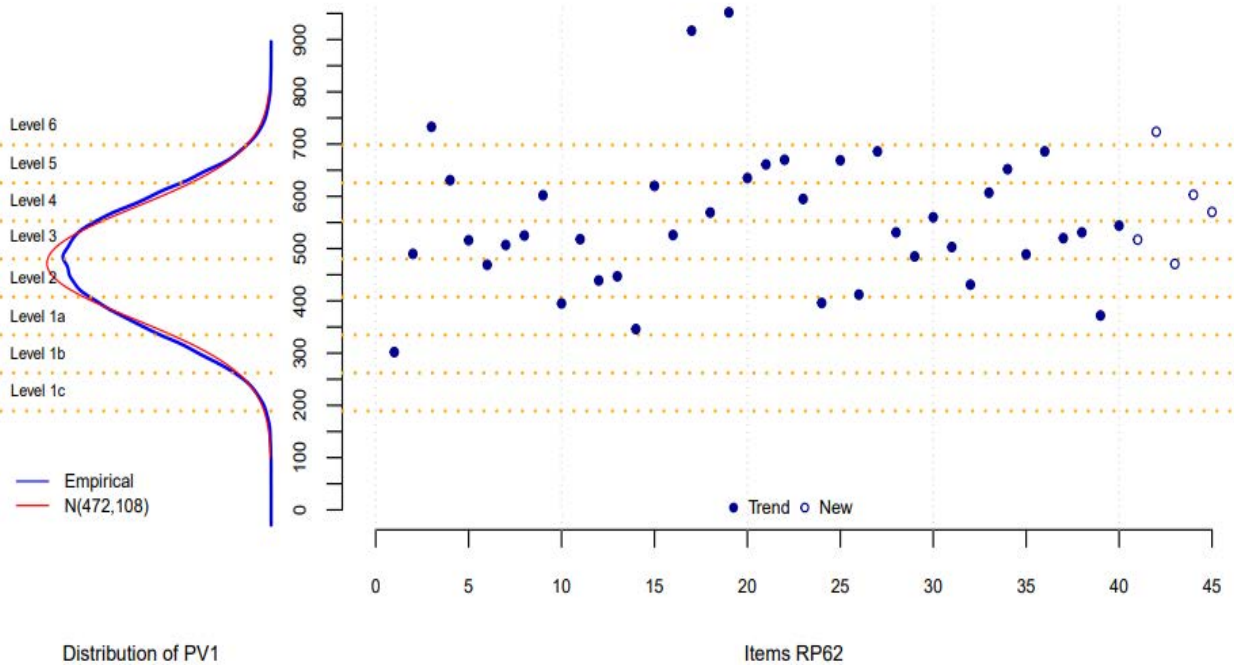


Figure 14.10. Distribution of the first plausible values and item RP62 values in financial literacy



### Population modelling outcomes

The population modelling outcomes include the multivariate latent regressions models estimated for each country/economy and the plausible values (PVs) generated from them, which are included in the international and national databases. Because the latent part of the population model comes from the IRT scaling, the plausible values are generated on their underlying PISA IRT metric used when estimating IRT

item and group parameters and then transformed to the PISA scale. For example, mathematics IRT scaled PVs are produced and then transformed to the PISA metric of mean 500 and standard deviation 100 across all participating OECD countries during the first mathematics assessment. Based on these PVs, then the overall and country/economy-level PISA scale reliability, average performance and students percentile by proficiency levels, and finally the correlations between domain scales were estimated. In the next sections, the methods used to transform the PVs from the IRT scale to the PISA reporting scales are described and the outcomes are reported.

### ***Mathematics, reading, science and financial literacy scaling transformations***

The mathematics, reading, and science PISA reporting scales were set when the domain became a major domain for the first time—in 2006 for science, 2009 for reading and 2012 in mathematics. This was done using a linear transformations of the senate weighted OECD participating countries/economies IRT scaled plausible values available at the time, so that the overall mean was 500 and the standard deviation 100, resulting in nearly all reported plausible values being between 200 and 800. The same approach was used for each new innovative domain and for the optional financial literacy domain.

However, because the IRT models used for scaling were updated in 2015, a bridge study was completed as part of the 2015 scaling analyses to establish new IRT to reported PISA scale parameters. This did not change the scales or the scores reported prior to 2015, but the new transformations have been applied since. Detailed descriptions of the bridge study and its results are provided in the PISA 2015 technical report OECD (2017<sup>[2]</sup>), Chapter 12)

Annex Table 14.A.9 provides the PISA IRT theta to reported PISA proficiency scale linear transformation A and B coefficients for the core and financial literacy domains. Given any IRT scaled theta ( $\theta$ ) value (e.g., item difficulty, item step parameters, or student PV or proficiency), the transformed value on the PISA scale is  $A * \theta + B$ .

### ***Creative Thinking scaling transformation***

For the creative thinking innovative domain developed for the PISA 2022 main survey, it was found that the use of a non-linear transformation provided a more appropriate reporting scale. This was because the particular challenges in creating such an innovative measure resulted in a relatively small pool that did not provide much information towards the lower end of the scale. To best support scale interpretations, the creative thinking item pool IRT test characteristic curve transformation of the theta plausible values  $PV_\theta$  was applied to obtain the reported plausible values  $PV_{NC} = \sum_{i \in V_p} T_i(PV_\theta)$ , where  $V_p$  indicates the set of 32 creative thinking items and  $T_i(PV_\theta)$  is the expected score on item  $i$  as a function of  $PV_\theta$  and item  $i$ 's IRT parameters. In this way the reported creative thinking plausible values can be interpreted as the expected number correct on a hypothetical form made up of all the items in the creative thinking item pool, given the proficiency level that the plausible value represents.

### ***Reliability of the PISA scales***

As was done in prior PISA cycles, test reliability was estimated using the well known theoretical formula:  $1 - (\text{expected error variance}/\text{total variance})$ . In practice, the expected error variance is the weighted average of the students' posterior variance, computed as the variance of the 10 plausible values, which is an expression of the posterior measurement error. The total variance was computed using a resampling approach (Efron, 1982<sup>[3]</sup>), using each country/economy set of resampling weights.

Annex Table 14.A.10 presents the test reliability descriptive statistics across countries/economies for the cognitive domains and the mathematics subscales. The reliabilities for each country/economy are presented in Annex Table 14.A.11. Overall, we observe that in average test reliability is high for the core

and financial literacy domains (0.84 to 0.90) and a bit less for creative thinking (0.8), and that most countries/economies' reliability is close to the average. As expected, since the number of items is smaller than for the full mathematics instrument, the reliability of the mathematics subscales are much lower and more variable cross countries/economies.

Annex Table 14.A.12 shows the average transformed plausible values as well as the resampling-based standard errors for each country and domain.

### ***Domain inter-correlations***

Estimated correlations between the domains, based on the 10 reported plausible values and averaged across all countries and assessment modes, are presented in Note: Ukrainian regions (18 out of 27) administered the assessment.

\*Denotes a country/economy for which the financial literacy domain was not fully sampled across the population; it is not a nationally representative sample. Note: Ukrainian regions (18 out of 27) administered the assessment.

\*Denotes a country/economy for which the financial literacy domain was not fully sampled across the population; it is not a nationally-representative sample.

Annex Table 14.A.13 and Annex Table 14.A.14 for the core domains and creative thinking, and in Annex Table 14.A.15 for the financial literacy sample. The estimated correlations for each country are presented in Annex Table 14.A.16.

### ***Mathematics subscales correlations***

There were two sets of subscales reported for mathematics. The first set, measuring content domains, was composed of the following four subscales: space and shape (MCSS), quantity (MCQN), change and relationships (MCCR), and uncertainty and data (MCUD). The second set, based on the cognitive processes, comprised the following four subscales: employing mathematical concepts, facts, and procedures (MPEM), interpreting, applying, and evaluating mathematical outcomes (MPIN), formulating situations mathematically (MPFS), and reasoning (MPRE).

The correlations between reading, science and the mathematics content domain subscales are presented in Annex Table 14.A.17. Annex Table 14.A.18 shows the correlations between reading, science and the cognitive process domains.

Note that, as indicated in Chapter 11, because of the way in which these subscale plausible values were estimated, it is not appropriate to correlate the cognitive process subscales with the cognitive contents subscales, or any of the subscales with the overall mathematics proficiency.

### ***Countries/economies average proficiency and percentages of students at each proficiency level***

Figure 14.11, Figure 14.12, Figure 14.13, Figure 14.14 and Figure 14.15 show the average proficiency and percentages of students at each proficiency level across countries/economies for each domain.

### **Linking error**

The estimation of the linking error between two PISA cycles was accomplished by considering the differences between the reported country means from the previous PISA cycles and new estimates of these country means based on the new PISA cycle item parameters. To estimate the linking error for trend

comparisons between PISA 2022 and a previous PISA cycle down to 2006, the subset of countries that had participated in both cycles being compared was used. In the cases of trends to 2000 or 2006 or financial literacy, since the number of participating countries was relatively small, all countries were used.

The 2022 linking errors are reported in Annex Table 14.A.19. Using these values help evaluate the extent to which changes in a country/economy or subgroup's performance between PISA 2022 and a previous PISA cycle are significantly different.

Note that for each domain, the earliest cycle for which comparisons can be made between PISA 2022 and a previous PISA cycle is the cycle in which the domain first became a major domain. Thus, the comparison of mathematics scores between PISA 2022 and PISA 2000 is not possible, nor is the comparison of science scores between PISA 2022 and PISA 2000 or between PISA 2022 and PISA 2003. Detail on the methodology used to calculate the linking errors can be found in Chapter 11.

**Table 14.1. Global Analysis of Item Dynamics and Student Proficiency in PISA 2022**

Figure	Title
Figure 14.1	Frequency of invariant, variant, and dropped items for mathematics, by country/economy
Figure 14.2	Frequency of invariant, variant, and dropped items for reading, by country/economy
Figure 14.3	Frequency of invariant, variant, and dropped items for science, by country/economy
Figure 14.4	Frequency of invariant, variant, and dropped items for creative thinking competence, by country/economy
Figure 14.5	Frequency of invariant, variant, and dropped items for financial literacy, by country/economy
Figure 14.6	Distribution of the first plausible values and item RP62 values in mathematics
Figure 14.7	Distribution of the first plausible values and item RP62 values in reading
Figure 14.8	Distribution of the first plausible values and item RP62 values in science
Figure 14.9	Distribution of the first plausible values and item RP62 values in creative thinking
Figure 14.10	Distribution of the first plausible values and item RP62 values in financial literacy
Figure 14.11	Percentage of students in each country/economy at each proficiency level for mathematics
Figure 14.12	Percentage of students in each country/economy at each proficiency level for reading
Figure 14.13	Percentage of students in each country/economy at each proficiency level for science
Figure 14.14	Percentage of students in each country/economy at each proficiency level for creative thinking
Figure 14.15	Percentage of students in each country/economy at each proficiency level for financial literacy

StatLink  <https://stat.link/4trwd2>

## References

- Efron, B. (1982), "The jackknife, the bootstrap, and other resampling plans", *CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38*, <https://doi.org/10.1137/1.9781611970319>. [3]
- OECD (2022), *PISA 2018 Technical Report*. [1]
- OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, <http://www.oecd.org/pisa/data/2015-technical-report/>. [2]

## Notes


- 
1. The dropped items are: CMA112Q02, CR547Q07S, DF082Q01C, and DT520Q01C, DT560Q01C, DT560Q02C, DT450Q01C, DT450Q02C and DT450Q03C.
  2. For the trend items classified as variant in a specific group (yellow), the 2018 parameters did not appropriately fit the 2022 data; thus, new unique parameters were estimated. For new items classified as variant in a specific group (yellow), unique parameters were needed due to the misfit of the common international parameters to the 2022 data.



# Annex 14.A. IRT Scaling Outcomes and Population Modelling Analysis

**Annex Table 14.A.1. Comprehensive Statistical Insights from PISA 2022: Item Dynamics, Proficiency Assessments, and Domain Interrelations**

Tables	Title
Table 14.A.2	Proportion of invariant, variant, and dropped CBA items averaged across countries/economies, for each domain
Table 14.A.3	Proportion of invariant, variant, and dropped new PBA items averaged across countries/economies, for each domain
Table 14.A.4	Proficiency levels for mathematics and the classification of items and students
Table 14.A.5	Proficiency levels for reading and the classification of items and students
Table 14.A.6	Proficiency levels for science and the classification of items and students
Table 14.A.7	Proficiency levels for creative thinking and the classification of items and students
Table 14.A.8	Proficiency levels for financial literacy and the classification of items and students
Table 14.A.9	PISA IRT theta to reported PISA proficiency scale linear transformation coefficients
Table 14.A.10	Test reliability descriptive statistics across countries/economies for the cognitive domains and the mathematics subscales
Table 14.A.11	Countries/economies reliability values for the cognitive domains
Table 14.A.12	Average plausible values (PV) and resampling-based standard errors (SE) by country and domain.
Table 14.A.13	Core domain inter-correlations for the main sample
Table 14.A.14	Creative Thinking inter-correlations with core domains for the main sample
Table 14.A.15	Domain inter-correlations for the financial literacy sample
Table 14.A.16	Domain inter-correlations by country/economy
Table 14.A.17	Mathematics content subscales inter-correlations
Table 14.A.18	Mathematics cognitive process subscales inter-correlations
Table 14.A.19	Linking error for score comparisons between PISA 2022 and previous PISA cycles

StatLink  <https://stat.link/y316ba>

**Annex Table 14.A.2. Proportion of invariant, variant, and dropped CBA items averaged across countries/economies, for each domain**

	Mathematics		Reading		Science	Financial	Creative
	Trend	New	Fluency	MSAT	All	Literacy	Thinking
Total items	74	159	65	196	115	40	32
Total countries	68	68	68	68	68	19	55
Invariant	86.0%	92.6%	93.7%	76.4%	83.0%	84.1%	77.4%
Group-specific invariant	6.5%	-	3.5%	11.8%	11.8%	2.6%	-
Invariant total <sup>1</sup>	92.5%	92.6%	97.2%	88.2%	94.8%	86.7%	77.4%
Noninvariant	6.1%	7.0%	1.7%	10.6%	4.0%	11.0%	20.8%
Dropped	1.4%	0.4%	1.1%	1.2%	1.2%	2.3%	1.8%

**Annex Table 14.A.3. Proportion of invariant, variant, and dropped new PBA items averaged across countries/economies, for each domain**

	Mathematics	Reading		Science
		Fluency	Reading	
Total items	64	79	66	66
Total countries	3	3	3	3
Invariant	82.3%	90.7%	79.3%	80.3%
Noninvariant	16.1%	9.3%	17.2%	18.7%
Dropped	1.6%	0.0%	3.5%	1.0%

**Annex Table 14.A.4. Proficiency levels for mathematics and the classification of items and students**

Classification	Number of items		Percentage of items		Percentage of respondents	
	CBA	New PBA	CBA	New PBA	CBA	New PBA
Level 6	38	1	16%	2%	2%	
Level 5	34	4	15%	6%	5%	0%
Level 4	50	7	21%	11%	10%	0%
Level 3	48	10	21%	16%	16%	3%
Level 2	45	25	19%	39%	21%	10%
Level 1a	13	8	6%	13%	23%	26%
Level 1b	5	6	2%	9%	24%	33%
Level 1c		2			3%	20%
Below Level 1		1		2%		7%

**Annex Table 14.A.5. Proficiency levels for reading and the classification of items and students**

Classification	Number of items				Percentage of items				Percentage of respondents	
	RF CBA	CBA	RF New PBA	New PBA	RF CBA	CBA	RF New PBA	New PBA	CBA	New PBA
Level 6		7		4		4%		6%	1%	
Level 5		14		1		7%		2%	4%	0%
Level 4		25		10		13%		15%	11%	1%
Level 3		42		10		21%		15%	19%	5%
Level 2		59	1	15		30%	1%	23%	24%	18%
Level 1a	1	40	3	18	2%	20%	4%	27%	23%	37%

Level 1b	22	8	30	3	34%	4%	38%	5%	15%	30%
Level 1c	11	1	46	5	17%	1%	58%	8%	5%	8%
Below Level 1	31				48%				1%	1%

**Annex Table 14.A.6. Proficiency levels for science and the classification of items and students**

Classification	Number of items		Percentage of items		Percentage of respondents	
	CBA	New PBA	CBA	New PBA	CBA	New PBA
Level 6	3	3	3%	4%	1%	
Level 5	15	8	13%	9%	4%	0%
Level 4	31	23	27%	27%	11%	0%
Level 3	36	30	31%	35%	20%	4%
Level 2	22	17	19%	20%	25%	18%
Level 1a	7	3	6%	4%	24%	43%
Level 1b	1	1	1%	1%	13%	30%
Below 1b					2%	5%

**Annex Table 14.A.7. Proficiency levels for creative thinking and the classification of items and students**

Classification	Number of items		Percentage of items		Percentage of respondents
	Partial Credit*	Full Credit	Partial Credit*	Full Credit	
Level 6		7		22%	4%
Level 5	2	17	7%	53%	16%
Level 4	7	6	25%	19%	19%
Level 3	11	2	39%	6%	22%
Level 2	6		21%		20%
Level 1	2		7%		16%
Below Level 1					4%

**Annex Table 14.A.8. Proficiency levels for financial literacy and the classification of items and students**

Classification	Number of items	Percentage of items	Percentage of respondents
Level 5	12	27%	8%
Level 4	8	18%	17%
Level 3	14	31%	25%
Level 2	6	13%	24%
Level 1	4	9%	17%
Below 1	1	2%	9%

**Annex Table 14.A.9. PISA IRT theta to reported PISA proficiency scale linear transformation coefficients**

Domain	A	B
Mathematics	135.9030	514.1848
Reading	131.5532	437.9244
Science	168.3189	494.5360
Financial literacy	140.0807	490.7259

Creative thinking*	-	-
--------------------	---	---

Note: \* Not applicable because a non-linear test characteristic curve transformation was used.

**Annex Table 14.A.10. Test reliability descriptive statistics across countries/economies for the cognitive domains and the mathematics subscales**

MODE	Domains	Median	S.D.	Max	Min	
CBA	Mathematics	0.90	0.03	0.93	0.81	
	Content	Change and Relationships	0.85	0.05	0.91	0.66
		Quantity	0.87	0.04	0.91	0.75
		Space and Shape	0.80	0.08	0.87	0.57
		Uncertainty and Data	0.84	0.05	0.90	0.71
	Cog. Process	Employing Mathematical Concepts, Facts, and Procedures	0.87	0.04	0.91	0.75
		Formulating Situations Mathematically	0.83	0.08	0.90	0.57
		Interpreting, Applying, and Evaluating Mathematical Outcomes	0.86	0.04	0.90	0.74
		Reasoning	0.85	0.08	0.91	0.59
	Reading	0.86	0.03	0.91	0.77	
	Science	0.87	0.03	0.92	0.79	
	Financial literacy	0.90	0.02	0.92	0.85	
	Creating Thinking	0.80	0.04	0.89	0.65	
PBA	Reading	0.87	0.03	0.90	0.84	
	Mathematics	0.87	0.01	0.89	0.85	
	Science	0.84	0.03	0.87	0.81	

**Annex Table 14.A.11. Countries/economies reliability values for the cognitive domains**

Mode	Country/Economy	Mathematics	Reading	Science	Financial Literacy	Creative Thinking
CBA	Albania	0.85	0.77	0.80		0.80
CBA	United Arab Emirates	0.90	0.86	0.85	0.89	0.77
CBA	Argentina	0.87	0.85	0.85		
CBA	Australia	0.92	0.85	0.87		0.76
CBA	Austria	0.93	0.90	0.91	0.91	
CBA	Belgium*	0.92	0.86	0.89	0.90	0.79
CBA	Bulgaria	0.90	0.87	0.86	0.89	0.82
CBA	Brazil	0.86	0.84	0.85	0.87	0.77
CBA	Brunei Darussalam	0.92	0.91	0.91		0.87
CBA	Canada*	0.89	0.82	0.83	0.90	0.69
CBA	Switzerland	0.92	0.90	0.91		
CBA	Chile	0.87	0.84	0.86		0.77
CBA	Colombia	0.87	0.85	0.86		0.81
CBA	Costa Rica	0.86	0.84	0.83	0.87	0.82
CBA	Czech Republic	0.92	0.87	0.89	0.90	0.79
CBA	Germany	0.92	0.88	0.90		0.82
CBA	Denmark	0.90	0.85	0.89	0.90	0.78
CBA	Dominican Republic	0.82	0.85	0.81		0.79
CBA	Spain	0.89	0.82	0.83	0.86	0.65
CBA	Estonia	0.90	0.84	0.86		0.77
CBA	Finland	0.91	0.86	0.87		0.81
CBA	France	0.92	0.87	0.88		0.80
CBA	United Kingdom	0.92	0.87	0.89		
CBA	Georgia	0.88	0.84	0.83		
CBA	Greece	0.89	0.84	0.86		0.81
CBA	Hong Kong (China)	0.92	0.85	0.86		0.77

Mode	Country/Economy	Mathematics	Reading	Science	Financial Literacy	Creative Thinking
CBA	Croatia	0.91	0.84	0.87		0.77
CBA	Hungary	0.92	0.88	0.90	0.90	0.84
CBA	Indonesia	0.85	0.87	0.85		0.81
CBA	Ireland	0.91	0.88	0.89		
CBA	Iceland	0.89	0.84	0.86		0.77
CBA	Israel	0.92	0.86	0.88		0.85
CBA	Italy	0.91	0.86	0.88	0.90	0.80
CBA	Jamaica	0.88	0.89	0.88		0.89
CBA	Jordan	0.81	0.81	0.82		0.79
CBA	Japan	0.92	0.86	0.89		
CBA	Kazakhstan	0.85	0.83	0.81		0.74
CBA	Korea	0.92	0.85	0.88		0.80
CBA	Kosovo	0.85	0.85	0.84		
CBA	Lithuania	0.91	0.85	0.89		0.81
CBA	Latvia	0.90	0.85	0.88		0.74
CBA	Macao (China)	0.91	0.84	0.88		0.80
CBA	Morocco	0.84	0.82	0.81		0.83
CBA	Republic of Moldova	0.89	0.88	0.86		0.81
CBA	Mexico	0.86	0.86	0.86		0.79
CBA	North Macedonia	0.88	0.84	0.84		0.85
CBA	Malta	0.91	0.87	0.88		0.85
CBA	Montenegro	0.89	0.86	0.86		
CBA	Mongolia	0.89	0.84	0.86		0.79
CBA	Malaysia	0.90	0.88	0.88	0.92	0.85
CBA	Netherlands	0.93	0.89	0.91	0.91	0.84
CBA	Norway	0.91	0.86	0.87	0.85	
CBA	New Zealand	0.92	0.88	0.89		0.83
CBA	Panama	0.87	0.88	0.88		0.85
CBA	Peru	0.87	0.84	0.86	0.89	0.80
CBA	Philippines	0.88	0.90	0.87		0.89
CBA	Poland	0.91	0.87	0.87	0.90	0.78
CBA	Portugal	0.91	0.85	0.88	0.87	0.78
CBA	Palestinian Authority	0.83	0.82	0.81		0.81
CBA	Qatar	0.91	0.87	0.88		0.84
CBA	Baku (Azerbaijan)	0.87	0.81	0.81		0.73
CBA	Cyprus	0.90	0.84	0.84		0.81
CBA	Ukrainian regions (18 of 27)	0.89	0.86	0.87		0.82
CBA	Romania	0.92	0.90	0.90		0.85
CBA	Saudi Arabia	0.83	0.81	0.79	0.86	0.76
CBA	Singapore	0.92	0.86	0.88		0.78
CBA	El Salvador	0.82	0.84	0.84		0.79
CBA	Serbia	0.90	0.86	0.87		0.79
CBA	Slovak Republic	0.92	0.87	0.89		0.85
CBA	Slovenia	0.91	0.87	0.90		0.82
CBA	Sweden	0.92	0.88	0.90		
CBA	Chinese Taipei	0.93	0.89	0.90		0.81
CBA	Thailand	0.88	0.86	0.87		0.85
CBA	Türkiye	0.92	0.88	0.90		
CBA	Uruguay	0.89	0.85	0.87		0.81
CBA	United States	0.92	0.90	0.92	0.91	
CBA	Uzbekistan	0.81	0.80	0.79		0.76
New PBA	Guatemala	0.88	0.90	0.87		
New PBA	Cambodia	0.85	0.84	0.81		
New PBA	Paraguay	0.89	0.89	0.87		

Mode	Country/Economy	Mathematics	Reading	Science	Financial Literacy	Creative Thinking
PBA	Viet Nam	0.87	0.84	0.81		

Note: Ukrainian regions (18 out of 27) administered the assessment.

\*Denotes a country/economy for which the financial literacy domain was not fully sampled across the population; it is not a nationally-representative sample.

**Annex Table 14.A.12 Average plausible values (PV) and resampling-based standard errors (SE) by country and domain**

Country	Reading		Mathematics		Science		Financial Literacy		Creative Thinking	
	Average PV	SE	Average PV	SE	Average PV	SE	Average PV	SE	Average PV	SE
International average	435.04	0.30	437.63	0.27	446.89	0.28	474.59	0.67	27.83	0.04
Albania	358.43	1.93	368.22	2.09	375.97	2.22			13.09	0.28
Argentina	400.74	2.57	377.53	2.25	406.19	2.49				
Australia	498.05	2.01	487.08	1.78	507.00	1.93			37.31	0.25
Austria	480.41	2.67	487.27	2.34	491.27	2.65	506.22	2.79		
Baku (Azerbaijan)	365.21	2.45	396.88	2.38	380.14	2.21			22.78	0.31
Belgium*	478.85	2.52	489.49	2.20	490.58	2.48	526.63	3.19	34.91	0.27
Brazil	410.36	2.09	378.69	1.58	403.00	1.93	415.51	2.29	23.32	0.29
Brunei Darussalam	429.23	1.16	442.09	0.93	445.86	1.32			23.74	0.19
Bulgaria	404.30	3.40	417.30	3.30	420.99	3.17	426.07	3.70	20.72	0.38
Cambodia	328.84	2.08	336.40	2.69	347.10	2.10				
Canada*	507.13	1.97	496.95	1.56	515.02	1.93	518.74	2.42	37.93	0.22
Chile	447.98	2.63	411.70	2.08	443.54	2.47			30.67	0.31
Chinese Taipei	515.17	3.25	547.09	3.78	537.38	3.31			32.62	0.39
Colombia	408.67	3.75	382.70	3.03	411.12	3.28			25.55	0.49
Costa Rica	415.23	2.66	384.58	1.89	410.99	2.42	418.23	3.10	27.48	0.32
Croatia	475.50	2.44	463.11	2.38	482.67	2.40			30.46	0.31
Cyprus	381.08	1.16	418.31	1.18	410.90	1.46			23.73	0.20
Czech Republic	488.60	2.25	487.00	2.09	497.74	2.30	506.61	2.22	32.64	0.29
Denmark	488.80	2.58	489.27	1.95	493.82	2.50	520.54	2.44	35.49	0.24
Dominican Republic	351.31	2.44	339.11	1.62	360.43	2.04			15.49	0.26
El Salvador	364.90	2.80	343.47	2.00	373.14	2.62			22.97	0.35
Estonia	511.03	2.36	509.95	1.98	525.81	2.07			35.85	0.27
Finland	490.22	2.26	484.14	1.86	510.96	2.50			35.82	0.30
France	473.85	3.07	473.94	2.49	487.23	2.73			32.43	0.31
Georgia	373.86	2.29	390.02	2.37	384.07	2.31				
Germany	479.79	3.61	474.83	3.06	492.43	3.48			32.53	0.40
Greece	438.44	2.83	430.15	2.34	440.79	2.77			27.00	0.33
Guatemala	374.12	2.44	344.20	2.21	372.96	2.23				
Hong Kong (China)	499.70	2.85	540.35	2.99	520.42	2.79			31.57	0.35
Hungary	472.97	2.83	472.78	2.51	485.89	2.71	492.41	3.11	30.94	0.33
Iceland	435.90	2.06	458.90	1.58	446.93	1.76			30.46	0.25
Indonesia	358.57	2.91	365.53	2.35	382.86	2.56			18.96	0.39
Ireland	516.01	2.33	491.65	2.02	503.85	2.26				
Israel	473.83	3.49	457.90	3.27	464.75	3.38			32.28	0.39
Italy	481.60	2.68	471.26	3.09	477.46	3.18	483.51	3.11	31.40	0.31
Jamaica	409.63	4.21	377.42	3.14	402.93	3.88			25.54	0.54
Japan	515.85	3.18	535.58	2.93	546.63	2.80				
Jordan	342.17	2.40	361.23	2.03	374.53	2.35			20.21	0.36
Kazakhstan	386.28	1.66	425.44	1.69	423.17	1.72			23.84	0.29
Korea	515.42	3.63	527.30	3.86	527.82	3.58			38.09	0.39

Country	Reading		Mathematics		Science		Financial Literacy		Creative Thinking	
	Average PV	SE	Average PV	SE	Average PV	SE	Average PV	SE	Average PV	SE
Kosovo	342.19	1.06	354.96	1.02	357.02	1.26				
Latvia	474.57	2.46	483.16	2.03	493.84	2.30			35.07	0.27
Lithuania	471.83	2.21	475.15	1.84	484.46	2.33			32.86	0.28
Macao (China)	510.41	1.35	551.92	1.10	543.10	1.11			31.62	0.20
Malaysia	388.09	2.75	408.69	2.40	416.31	2.35	405.75	2.94	25.11	0.38
Malta	445.30	1.90	466.02	1.58	465.59	1.70			31.32	0.24
Mexico	415.36	2.92	395.03	2.27	409.89	2.42			28.99	0.32
Mongolia	378.42	2.25	424.59	2.57	412.38	2.36			24.92	0.32
Montenegro	405.02	1.35	405.60	1.12	403.13	1.21				
Morocco	339.36	3.97	364.77	3.35	365.40	3.38			15.48	0.58
Netherlands	459.24	4.28	492.68	3.77	488.32	4.07	516.88	4.42	32.39	0.46
New Zealand	500.85	2.12	479.07	1.99	504.13	2.24			36.43	0.29
North Macedonia	358.52	0.81	388.58	0.87	379.88	0.93			19.11	0.23
Norway	476.52	2.54	468.45	2.06	478.23	2.37	488.73	2.63		
Palestinian Authority	349.16	2.03	365.75	1.84	368.82	2.10			18.46	0.34
Panama	391.95	3.41	356.57	2.84	387.77	3.54			23.22	0.34
Paraguay	373.16	2.44	337.54	2.16	368.33	2.06				
Peru	408.25	2.73	391.24	2.34	407.78	2.64	420.75	3.04	23.45	0.35
Philippines	346.55	3.40	354.72	2.58	356.17	3.11			14.20	0.51
Poland	488.71	2.74	488.96	2.27	499.16	2.55	505.84	2.69	34.44	0.28
Portugal	476.59	2.66	471.91	2.35	484.37	2.56	494.42	2.37	33.90	0.29
Qatar	419.30	1.45	414.11	1.14	432.40	1.48			27.66	0.24
Republic of Moldova	410.94	2.51	414.20	2.31	416.86	2.39			23.95	0.32
Romania	428.50	3.98	427.76	4.00	427.51	3.87			26.25	0.47
Saudi Arabia	382.55	1.99	388.78	1.76	390.39	1.96	412.47	2.58	23.32	0.31
Serbia	440.35	2.79	439.88	2.97	447.46	2.89			28.68	0.35
Singapore	542.55	1.87	574.66	1.23	561.43	1.33			40.96	0.17
Slovak Republic	446.86	3.10	463.99	2.89	462.27	3.03			29.22	0.40
Slovenia	468.54	1.64	484.53	1.24	499.96	1.45			29.99	0.23
Spain	474.31	1.65	473.14	1.50	484.53	1.60	486.08	2.70	32.75	0.22
Sweden	486.98	2.49	481.77	2.06	493.55	2.35				
Switzerland	483.33	2.26	507.99	2.14	502.52	2.19				
Thailand	378.66	2.82	393.95	2.68	409.26	2.78			20.93	0.37
Türkiye	456.08	1.85	453.15	1.59	475.94	1.93				
Ukrainian regions (18 of 27)	427.53	3.93	440.85	4.06	450.19	3.78			26.89	0.61
United Arab Emirates	417.35	1.34	431.11	0.95	431.98	1.31	441.12	1.58	28.43	0.16
United Kingdom	494.40	2.37	488.98	2.22	499.67	2.38				
United States	503.94	4.33	464.89	4.01	499.41	4.32	505.23	4.92		
Uruguay	430.36	2.41	408.71	2.02	435.38	2.48			28.64	0.35
Uzbekistan	335.50	2.00	363.94	2.02	354.86	2.01			14.50	0.25
Viet Nam	461.89	3.94	469.40	3.93	472.38	3.59				

Note: Ukrainian regions (18 out of 27) administered the assessment.

\*Denotes a country/economy for which the financial literacy domain was not fully sampled across the population; it is not a nationally-representative sample.

### Annex Table 14.A.13. Core domain inter-correlations for the main sample

DOMAIN	Reading	Science
Mathematics	Average	<b>0.80</b>
	Average (CBA)	0.80
	Average (PBA)	0.81

	Range	0.65 ~ 0.89	0.75 ~ 0.92
Reading	Average		<b>0.79</b>
	Average (CBA)		0.79
	Average (PBA)		0.81
	Range		0.67~ 0.88

Annex Table 14.A.14. Creative Thinking inter-correlations with core domains for the main sample

DOMAIN		Mathematics	Reading	Science
Creative Thinking	Average	<b>0.68</b>	<b>0.68</b>	<b>0.67</b>
	Range	0.53 ~ 0.80	0.55 ~ 0.83	0.54 ~ 0.80

Annex Table 14.A.15. Domain inter-correlations for the financial literacy sample

DOMAIN		Mathematics	Reading
Financial Literacy	Average	<b>0.86</b>	<b>0.84</b>
	Range	0.80 ~ 0.90	0.79 ~ 0.88

Annex Table 14.A.16. Domain inter-correlations by country/economy

Country	Mathematics & Reading	Mathematics & Science	Mathematics & Financial literacy	Mathematics & Creative Thinking	Reading & Science	Reading & Financial literacy	Reading & Creative Thinking	Science & Creative Thinking
Albania	0.69	0.76		0.66	0.67		0.58	0.60
Argentina	0.75	0.81			0.75			
Australia	0.80	0.86		0.65	0.78		0.63	0.64
Austria	0.84	0.90	0.88		0.85	0.86		
Baku (Azerbaijan)	0.75	0.82		0.64	0.72		0.63	0.63
Belgium*	0.82	0.90	0.89	0.69	0.82	0.85	0.69	0.69
Brazil	0.80	0.84	0.84	0.69	0.78	0.82	0.70	0.68
Brunei Darussalam	0.88	0.92		0.80	0.87		0.81	0.80
Bulgaria	0.83	0.86	0.87	0.76	0.80	0.85	0.74	0.74
Cambodia	0.79	0.78			0.75			
Canada*	0.75	0.80	0.85	0.56	0.72	0.81	0.55	0.54
Chile	0.79	0.86		0.61	0.78		0.58	0.57
Chinese Taipei	0.84	0.90		0.68	0.82		0.67	0.67
Colombia	0.80	0.86		0.69	0.77		0.68	0.68
Costa Rica	0.79	0.83	0.87	0.71	0.78	0.83	0.69	0.66
Croatia	0.79	0.86		0.67	0.77		0.67	0.68
Cyprus	0.75	0.82		0.72	0.74		0.70	0.68
Czech Republic	0.81	0.88	0.87	0.68	0.80	0.83	0.67	0.68
Denmark	0.79	0.87	0.87	0.62	0.77	0.84	0.61	0.61
Dominican Republic	0.80	0.80		0.64	0.77		0.67	0.62
El Salvador	0.80	0.81		0.67	0.76		0.66	0.66
Estonia	0.77	0.86		0.62	0.74		0.58	0.62
Finland	0.79	0.87		0.68	0.77		0.71	0.70
France	0.84	0.88		0.71	0.82		0.72	0.70
Georgia	0.75	0.81			0.74			
Germany	0.85	0.90		0.76	0.86		0.76	0.76
Greece	0.78	0.83		0.69	0.77		0.65	0.68



Country	Mathematics & Reading	Mathematics & Science	Mathematics & Financial literacy	Mathematics & Creative Thinking	Reading & Science	Reading & Financial literacy	Reading & Creative Thinking	Science & Creative Thinking
Guatemala	0.84	0.87			0.88			
Hong Kong (China)	0.79	0.84		0.63	0.76		0.61	0.60
Hungary	0.84	0.91	0.89	0.76	0.84	0.85	0.74	0.74
Iceland	0.77	0.85		0.67	0.77		0.68	0.68
Indonesia	0.78	0.77		0.57	0.72		0.55	0.54
Ireland	0.81	0.88			0.84			
Israel	0.81	0.88		0.76	0.80		0.74	0.73
Italy	0.76	0.84	0.82	0.64	0.75	0.79	0.62	0.61
Jamaica	0.84	0.86		0.67	0.82		0.71	0.69
Japan	0.81	0.88			0.84			
Jordan	0.72	0.80		0.66	0.74		0.68	0.66
Kazakhstan	0.65	0.75		0.53	0.71		0.62	0.59
Korea	0.76	0.85		0.59	0.74		0.59	0.61
Kosovo	0.78	0.83			0.77			
Latvia	0.79	0.88		0.57	0.78		0.55	0.57
Lithuania	0.81	0.88		0.71	0.81		0.69	0.69
Macao (China)	0.75	0.87		0.66	0.78		0.64	0.66
Malaysia	0.79	0.87	0.90	0.75	0.82	0.88	0.79	0.78
Malta	0.78	0.87		0.73	0.80		0.73	0.72
Mexico	0.82	0.86		0.66	0.80		0.67	0.66
Mongolia	0.79	0.87		0.71	0.78		0.69	0.70
Montenegro	0.79	0.86			0.77			
Morocco	0.77	0.83		0.72	0.75		0.70	0.68
Netherlands	0.86	0.90	0.90	0.72	0.85	0.88	0.74	0.71
New Zealand	0.81	0.88		0.69	0.85		0.71	0.71
North Macedonia	0.80	0.84		0.75	0.76		0.72	0.74
Norway	0.78	0.86	0.80		0.80	0.82		
Palestinian Authority	0.76	0.81		0.71	0.72		0.67	0.67
Panama	0.82	0.86		0.64	0.79		0.66	0.65
Paraguay	0.84	0.87			0.86			
Peru	0.82	0.86	0.86	0.71	0.79	0.87	0.70	0.68
Philippines	0.89	0.87		0.80	0.85		0.83	0.77
Poland	0.81	0.87	0.87	0.70	0.80	0.84	0.68	0.68
Portugal	0.81	0.87	0.85	0.70	0.80	0.83	0.70	0.69
Qatar	0.81	0.86		0.72	0.79		0.70	0.70
Republic of Moldova	0.83	0.87		0.70	0.81		0.74	0.71
Romania	0.86	0.90		0.78	0.85		0.77	0.77
Saudi Arabia	0.75	0.78	0.80	0.66	0.73	0.82	0.67	0.65
Serbia	0.81	0.87		0.70	0.79		0.68	0.70
Singapore	0.82	0.89		0.67	0.81		0.66	0.66
Slovak Republic	0.83	0.89		0.74	0.81		0.72	0.73
Slovenia	0.77	0.89		0.60	0.77		0.59	0.58
Spain	0.76	0.82	0.86	0.59	0.75	0.79	0.59	0.58
Sweden	0.81	0.88			0.84			
Switzerland	0.83	0.89			0.86			
Thailand	0.79	0.83		0.68	0.77		0.67	0.68
Türkiye	0.82	0.90			0.83			
Ukrainian regions (18 of 27)	0.79	0.86		0.72	0.79		0.67	0.72
United Arab Emirates	0.81	0.85	0.87	0.71	0.79	0.84	0.71	0.69
United Kingdom	0.81	0.86			0.79			
United States	0.83	0.89	0.89		0.87	0.86		
Uruguay	0.80	0.87		0.72	0.79		0.69	0.71

Country	Mathematics & Reading	Mathematics & Science	Mathematics & Financial literacy	Mathematics & Creative Thinking	Reading & Science	Reading & Financial literacy	Reading & Creative Thinking	Science & Creative Thinking
Uzbekistan	0.72	0.78		0.67	0.70		0.63	0.63
Viet Nam	0.77	0.82			0.75			

Note: Ukrainian regions (18 out of 27) administered the assessment.

\*Denotes a country/economy for which the financial literacy domain was not fully sampled across the population; it is not a nationally-representative sample.

#### Annex Table 14.A.17. Mathematics content subscales inter-correlations

	MCCR1	MCQN2	MCSS3	MCUD4
Reading	0.71	0.72	0.63	0.71
Science	0.76	0.77	0.68	0.75
MCCR1		0.86	0.77	0.82
MCQN2			0.79	0.85
MCSS3				0.76

#### Annex Table 14.A.18. Mathematics cognitive process subscales inter-correlations

	MPEM1	MPFS2	MPIN3	MPRE4
Reading	0.72	0.66	0.73	0.69
Science	0.77	0.71	0.77	0.74
MPEM1		0.83	0.87	0.84
MPFS2			0.81	0.79
MPIN3				0.82

#### Annex Table 14.A.19. Linking error for score comparisons between PISA 2022 and previous PISA cycles

Comparison	Mathematics	Reading	Science	Financial literacy
PISA 2000 to 2022		6.67		
PISA 2003 to 2022	5.55	5.25		
PISA 2006 to 2022	4.09	8.56	3.68	
PISA 2009 to 2022	4.28	4.66	5.92	
PISA 2012 to 2022	3.58	6.01	5.20	4.05
PISA 2015 to 2022	2.74	3.63	1.38	3.47
PISA 2018 to 2022	2.24	1.47	1.61	2.20

# 15 Coding Design, Coding Process, and Reliability Studies

## Introduction

The PISA 2022 assessment consisted of both constructed-response (CR) and multiple-choice items (MC). MC items could be simple multiple choice, with a single correct response selection, or complex multiple choice, with multiple correct response selections required. MC items had a predefined correct answer that could be computer coded. While a few CR items were designed to be coded by computer, most required a person to read the response and provide a code or score. These items are referred to as human-coded constructed response items.

This chapter describes the design, preparation, and processing of coding human-coded constructed-response (CR) items, and reports the reliability statistics and volume of responses that could be automatically coded for these items. A summary of all test items by domain, item format, and coding method is shown in Annex Table 15.A.2.

The CBA mathematics assessment was administered within each country/economy as both a linear test and a Multistage Adaptive Test (MSAT). The CBA reading assessment was also administered as an MSAT with three stages, while the science assessment was administered using a linear design. Countries participating in the CBA also had the option of administering a Financial Literacy assessment and a Creative Thinking assessment. One country chose to participate in the paper-based assessment (PBA), which has been administered since 2015, and three countries participated in the new paper-based assessment. More on the PISA 2022 test design is presented in Chapter 2 of this Technical Report.

## Coding design

Coding designs for CBA, PBA, and the new PBA were developed to accommodate the various needs of countries/economies in terms of the number of languages assessed, sample size, and assessed domains (i.e., meaning whether Financial Literacy or the innovative domain were to be coded in the country/economy). In general, it was expected that coders would be able to code approximately 1 000 responses per day, over a two- to three-week period. The number of expected student responses per domain was based on the sample size completing the assessment in each assessed language in the core domains and in the optional domains of Financial Literacy and Creative Thinking.

Annex Table 15.A.3 shows the number of coders recommended by domain in the CBA coding designs based on the sample size. This design is exclusive by language of assessment. CBA participants were able to determine the appropriate design for their country/economy and language(s) with a coding estimation tool, which estimated the coding workload for each coder (duration of coding and the number of responses to be coded by each coder).

Annex Table 15.A.3 also includes an example of this estimated workload.

PBA and new PBA countries' sample sizes had little variation, and there were no additional domain options; therefore, all countries participating in these assessments were advised to recruit six coders for each domain. Annex Table 15.A.4 shows the estimated workload for six coders in each domain.

### ***Designs for within-country and across-country scoring 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 (Shin, von Davier and Yamamoto, 2019<sup>[1]</sup>). Throughout the chapter, we use the term *coding* to refer to the assignment of a numerical value to a student text response, which indicates the type of response provided by the student, and the term *scoring* to refer to the assignment of full credit, partial credit, or no credit, which is derived from the codes applied. Scoring reliability in PISA 2022 was evaluated and reported at both within- and across-country levels.

The purpose of monitoring and evaluating within-country scoring reliability is to ensure accurate scoring of student responses across coders in the same county-by-language group and identify any coding inconsistencies or problems in the scoring process throughout the process so that they can be promptly addressed and resolved. Within-country scoring reliability was evaluated by reviewing the codes assigned by two or more human coders on the same student responses in a process called multiple coding. *Multiple coding* refers to the coding of the same student response data by different independent coders, such that inter-rater agreement statistics can be calculated and evaluated.

It was also important to check the consistency of coders across countries and language groups. Accurate and consistent *scoring* (full credit, partial credit, no credit) within a country does not necessarily mean that coders from all countries and language groups are applying the coding rubric in the same manner. Coding bias may be introduced if, for example, one country codes a certain type of response differently than other countries (Shin, von Davier and Yamamoto, 2019<sup>[1]</sup>). Across-country scoring reliability was evaluated by checking the correctness of the codes assigned by two bilingual human coders on a set of English anchor responses in a process called anchor coding. *Anchor coding* refers to the coding of a set of common (across-country-by-language groups) responses in English for each item, for which the correct code for each response is already known by the PISA international contractor (but not provided to coders). Because countries coded the same anchor responses for each human-coded CR item, their coding results on the anchor responses could be compared to the anchor key and, thereby, to each other. For each human-coded CR item, a set of thirty anchor responses in CBA, and ten in PBA and New PBA, were distributed to the designated bilingual coders for coding.

In CBA, item responses were randomly selected from all student responses and gathered into coding sets for multiple coding. In the domains of Mathematics, Science, Financial Literacy, and Creative Thinking, one coding set was compiled, such that all coders contributed to the multiple-coding agreement for all items. In the domain of Reading, items were distributed among four coding sets, such that each coder only saw responses to half of the items and thus contributed only to the scoring reliability for the items in their assigned coding set. Each domain had two bilingual coders – always coders 01 and 03 – who additionally coded thirty anchor responses in English for each item in their coding set. The design for multiple coding for the CBA is shown in Figure 15.1.

For multiple coding in the paper-based designs, student test booklets are first sorted by booklet number. Because each test booklet contains responses from two administered domains (for example, Mathematics and Science were administered in booklets 1-6 in PBA), coding sets are first multiple coded by coders in one administered domain and then single coded by the coders in the other administered domain. A specified number of booklets (52 booklets of each booklet number in PBA and 90 of each number in the new PBA) are designated for multiple coding. These booklets are distributed equally among six coding

sets and distributed to coders. In PBA, all coders code all coding sets, whereas a subset of coders code each coding set in new PBA. The PBA and new PBA coding designs are shown in Figure 15.2.

Figure 15.1. Organization of multiple coding for the CBA designs

Mathematics		Coder IDs															
		301 (bilingual)	302	303 (bilingual)	304	305	306	307	308	309	310	311	312	313	314	315	316
	Number of Coded Responses																
Coding Set	128/item	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Anchor Set	30/item	•	•														

Reading		Coder IDs																														
		201 (bilingual)	202	203 (bilingual)	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231
	Number of Coded Responses																															
Coding Set 1	100/item	•	•					•	•	•			•	•			•	•			•	•			•	•			•	•		
Coding Set 2		•	•			•	•			•			•			•			•			•			•			•			•	
Coding Set 3				•	•	•	•				•	•			•	•			•	•			•	•			•	•			•	
Coding Set 4				•	•			•	•			•			•			•			•			•			•			•		•
Anchor Set	30/item	•	•																													

Science		Coder IDs										
		101 (bilingual)	102	103 (bilingual)	104	105	106	107	108	109	110	111
	Number of Coded Responses											
Coding Set	128/item	•	•	•	•	•	•	•	•	•	•	•
Anchor Set	30/item	•	•									

Financial Literacy		Coder IDs										
		401 (bilingual)	402	403 (bilingual)	404	405	406	407	408	409	410	411
	Number of Coded Responses											
Coding Set	100/item	•	•	•	•	•	•	•	•	•	•	•
Anchor Set	30/item	•	•									

Creative Thinking		Coder IDs																							
		501 (bilingual)	502	503 (bilingual)	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524
	Number of Coded Responses																								
Coding Set	100/item	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Anchor Set	30/item	•	•																						

Figure 15.2. Organization of multiple coding for the PBA and New PBA standard coding design

PBA			Coder IDs					
			301 (bilingual)	302	303 (bilingual)	304	305	306
			Number of Coded Responses					
<b>Mathematics</b>	Booklets 7-12 52 of each booklet number (312 total)	Coding Set 13	•	•	•	•	•	•
		Coding Set 14	•	•	•	•	•	•
		Coding Set 15	•	•	•	•	•	•
		Coding Set 16	•	•	•	•	•	•
		Coding Set 17	•	•	•	•	•	•
		Coding Set 18	•	•	•	•	•	•
<b>Booklet 92</b>	Anchor Coding Set	10/item	•	•				
			Number of Coded Responses					
			201 (bilingual)	202	202 (bilingual)	204	205	206
<b>Reading</b>	Booklets 13-18 52 of each booklet number (312 total)	Coding Set 25	•	•	•	•	•	•
		Coding Set 26	•	•	•	•	•	•
		Coding Set 27	•	•	•	•	•	•
		Coding Set 28	•	•	•	•	•	•
		Coding Set 29	•	•	•	•	•	•
		Coding Set 30	•	•	•	•	•	•
<b>Booklet 91</b>	Anchor Coding Set	10/item	•	•				
			Number of Coded Responses					
			101 (bilingual)	102	102 (bilingual)	104	105	106
<b>Science</b>	Booklets 1-6 52 of each booklet number (312 total)	Coding Set 1	•	•	•	•	•	•
		Coding Set 2	•	•	•	•	•	•
		Coding Set 3	•	•	•	•	•	•
		Coding Set 4	•	•	•	•	•	•
		Coding Set 5	•	•	•	•	•	•
		Coding Set 6	•	•	•	•	•	•
<b>Booklet 93</b>	Anchor Coding Set	10/item	•	•				

New PBA			Coder IDs					
			301 (bilingual)	302	303 (bilingual)	304	305	306
			Number of Coded Responses					
<b>Mathematics</b>	Booklets 5-8 90 of each booklet number (360 total)	Coding Set 13	•	•	•	•	•	•
		Coding Set 14	•	•	•	•	•	•
		Coding Set 15	•	•	•	•	•	•
		Coding Set 16	•	•	•	•	•	•
		Coding Set 17	•	•	•	•	•	•
		Coding Set 18	•	•	•	•	•	•
<b>Booklet 93</b>	Anchor Coding Set	10/item	•	•				
			Number of Coded Responses					
			201 (bilingual)	202	202 (bilingual)	204	205	206
<b>Reading</b>	Booklets 1-4 90 of each booklet number (360 total)	Coding Set 1	•	•	•	•	•	•
		Coding Set 2	•	•	•	•	•	•
		Coding Set 3	•	•	•	•	•	•
		Coding Set 4	•	•	•	•	•	•
		Coding Set 5	•	•	•	•	•	•
		Coding Set 6	•	•	•	•	•	•
<b>Booklet 92</b>	Anchor Coding Set	10/item	•	•				
			Number of Coded Responses					
			101 (bilingual)	102	102 (bilingual)	104	105	106
<b>Science</b>	Booklets 9-12 90 of each booklet number (360 total)	Coding Set 25	•	•	•	•	•	•
		Coding Set 26	•	•	•	•	•	•
		Coding Set 27	•	•	•	•	•	•
		Coding Set 28	•	•	•	•	•	•
		Coding Set 29	•	•	•	•	•	•
		Coding Set 30	•	•	•	•	•	•
<b>Booklet 91</b>	Anchor Coding Set	10/item	•	•				

## Coding preparation

Prior to the assessment, key activities were completed by National Centres to prepare for the process of coding CR items.

### Recruitment of national coder teams

The first task of National Project Managers (NPMs) on the coding workflow was to assemble a national coder team. The size of the coding teams varied in each country, but the following criteria were used for selecting members of the team:

- All coders should have more than an upper secondary education qualification (i.e., high school degree); university graduates were preferred.
- All should have a good understanding of secondary education 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 in the language(s) of the assessment.

After the national coding team was assembled, the next task was to identify a *lead coder* who was part of the coding team but also responsible for the following tasks:

- training coders within the country,
- organising all materials and distributing them to coders,
- monitoring the coding process,
- monitoring inter-rater reliability and taking action when the coding results were unacceptable or required further investigation,
- Producing reliability reports
- retraining or replacing coders if necessary, and
- consulting with the international experts if item-specific issues arose.

Additionally, the lead coder was required to be proficient in English, as international trainings and interactions with the PISA international contractors were in English only, and was encouraged to attend the international coder trainings. 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 main survey.

### ***Coder training materials***

Detailed coding guides were developed for all the new items in the domains of Mathematics, Financial Literacy, and Creative Thinking. These coding guides included coding rubrics for each item and example responses corresponding to each level (i.e., correct, partially correct, and incorrect) of the rubric. Coding rubrics for new items were revised for the main survey based on information learned from the field trial. Coding guides for trend domains were also prepared, but changes were limited to the correction of errors.

In addition to the coding guides, a separate workshop-materials file was either created for new domains or updated for trend domains. Unlike the coding guides which remain relatively static across cycles, the workshop-materials file can be updated. The workshop materials files contain additional example responses and annotations, which could be used to supplement the coder trainings. The additional example responses better illustrate the depth and breadth of the coding levels, and the lines between levels. Following the international trainings, final versions of all materials were prepared and released to participating countries/economies.

### ***International coder trainings***

Prior to the field trial, NPMs and lead coders were provided with a full item-by-item coder training for CBA, PBA, and new PBA participants in Athens, Greece in January 2020. The field trial training covered all items in all domains. Due to the one-year delay caused by the COVID-19 pandemic, a second international field trial training was held in January and February 2021. The second field trial training took place over several sessions and was conducted virtually. Additionally, the second field trial training covered only new material. That is, the sessions offered were for the new Mathematics items, all Creative Thinking items, and the four new Financial Literacy CR items.

Prior to the main survey, international coder trainings were held in January and February 2022, and were again conducted virtually for all domains. Full trainings were offered for all the new and all the trend Mathematics items, all the Creative Thinking items, and the new Financial Literacy items. Targeted

trainings were offered for the trend domains (Science, Reading, and the trend items in Financial Literacy); that is, the international experts reviewed analysis results from the field trial and considered items that have been historically challenging to code, and targeted items for which a refresher training would be most beneficial. Participants were also given the opportunity to ask questions about trend items if they were not already on the list prepared by the experts. Participants were also provided with the recorded trainings that were prepared for PISA 2018, which cover the items in the trend domains, and could be used to supplement the targeted virtual trainings.

Full trainings were provided virtually in April and May 2022 for PBA and New PBA participants for all domains. During these trainings, the coding guides were presented and explained, and participants had the opportunity to ask questions to have the coding rubrics clarified. Participants also practiced coding on sample responses and discussed any ambiguous or problematic situations as a group. When the discussion revealed areas where rubrics could be improved, those changes were noted and eventually implemented in an updated version of the coding guide that was made available after the meeting. The workshop-materials files were also updated as needed following the international trainings.

To support the national teams during the coding process, a coding query service was offered, which allowed national teams to submit coding questions and receive responses from the relevant domain experts. National teams were also able to see questions submitted by other countries/economies pertaining to the coding of new items, 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 stored on the PISA 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 and replicate as much as possible from the international training for their own coders. The training package consisted of an overview of the survey and their own training manuals based on the source manuals and materials provided by the PISA international contractors. Coding teams were asked to facilitate discussion about any items that were challenging to code. Past experience has shown that when coders discuss items among themselves and with their lead coder, many issues can be resolved, and more consistent coding can be achieved.

The National Centres were responsible for organising training and coding. The recommended approach was to train at the item level. Under this approach, coders were fully trained on the coding rules for one item, and then proceeded with coding all responses for that one item. Once the item was fully coded, training was provided for the next item (blocked by unit), and so on. The approach of coding item by item has been shown to improve reliability by helping coders to apply the scoring rubric more consistently.

For PBA and new PBA participants, coder training was also recommended at the item level; however, training could be given at the unit level. Once the training was complete on the items within a single unit, coding could take place across booklet for all the items within that one unit.

## **Coding procedures**

Since PISA 2015, coding CBA item responses has been facilitated through use of the Open-Ended Coding System (OECS), which allows coders to view student responses, defer responses for further review, and code responses directly in the system interface. The OECS supported coding teams in their work to code the CBA responses while ensuring that the coding design was appropriately implemented. Especially important during the COVID-19 pandemic, the OECS afforded coders the ability to work remotely. Detailed information about the system was included in the OECS manual provided to countries/economies.



Computer-based responses were coded on an item-by-item basis. For each item, coders receive a set of responses to be coded. Each set includes 1) student responses to be multiple coded as part of the within-country reliability monitoring process, and 2) student responses to be single coded. If the coder is one of the national team's two bilingual coders they also received anchor responses in English will also be included for across-country reliability monitoring. Because the generation of inter-rater agreement statistics were continuously being updated by the OECS as coders code (see Formula 13.1 in the *Reliability Studies* section), no pause in coding is required in the CBA to manually calculate these statistics, allowing coders to work at their own pace through all assigned responses.

When a coder logs into the system and selects an item to code, responses that require human coding appear on screen. Buttons at the top of the screen allow the coder to scroll through responses. In general, multiple-coded responses are populated first and then single-coded responses; for bilingual coders, anchor responses appear ahead of all student responses. For each response, the OECS displays the item stem or question, the individual response, and the available codes for the item, as well as a checkbox to *defer* the response to the lead coder and a checkbox to indicate that the response has been *recoded* from the originally applied code to a new code for some reason. It is expected that coders will code most responses assigned to them and defer responses only in unusual circumstances. When deferring a response, coders were encouraged to note the reason for deferral into an associated comment box. Coders generally worked on one item at a time until all responses in that item set were coded. The process was repeated until responses for all items were coded. Detailed information about the system was provided in the OECS manual.

For the PBA and New PBA, the coding designs were supported by the Data Management Expert (DME) system, and reliability was monitored through the Open-Ended Reporting System (OERS), an additional software that worked in conjunction with the DME to evaluate and report reliability for CR items. The coding process for paper-based participants involved using the actual paper booklets, with sections of some booklets single-coded and some sections coded multiple times. When a response is single coded, coders mark directly in the booklets. When a response is coded multiple times, only the final coder codes directly in the booklet, while all others code on coding sheets; this allows coders to remain independent in their coding decisions and provides an accurate evaluation of scoring reliability. Detailed information about the system was provided in the OERS manual to PBA countries/economies.

Unlike coding in the CBA, the process of coding in PBA and New PBA does require a pause between coding different sets of responses (anchor responses, multiple-coded responses, and single-coded responses), resulting in three distinct coding phases. In the first phase of coding, bilingual coders code the anchor responses, enter the data into the project database using the DME and evaluate the across-country scoring reliability using the OERS. In the second phase, at least 100 student responses for each item are multiple coded. Single-coded responses are addressed in the final phase. All anchor- and multiple-coded response codes are entered into the project database using the DME and run the OERS reliability software for review. Any coding issues identified by the OERS are investigated and corrected before moving forward. The distributions of single codes are also reviewed in the OERS, as a quality check.

National Centres used the output reports generated by the OECS and OERS to monitor irregularities and deviations in the coding process. The OECS and OERS generate the following reports of scoring reliability: i) percentage of first-digit code agreement on multiple and anchor coded responses and ii) coding category distribution across coders. NPMs were instructed to investigate whether a systematic pattern of irregularities existed and if the observed pattern was attributable to a particular coder or item. In addition, NPMs were instructed not to carry out *coding resolution* (changing coding on individual responses to reach higher coding consistency). Instead, if systematic irregularities were identified, coders were to be retrained and all responses from a particular item or a particular coder were to be recoded, including those codes that showed agreement. Coding inconsistencies usually come from a misunderstanding of the general

coding guidelines and/or a rubric for a particular item. Reliability studies conducted by the PISA contractors also made use of the OECS and OERS reports submitted by National Centres.

## Reliability studies

Careful monitoring of scoring reliability plays an important role in data quality control. National Centres used the output reports generated by the OECS and OERS to monitor irregularities and deviations in the coding process for both items and individual coders. Through these processes of reliability monitoring, coding inconsistencies or problems within and across countries could be detected early in the coding process, and action could be taken quickly to address these concerns.

### ***Within-country monitoring of scoring reliability***

While coding was ongoing, score agreement and coding category distribution were the main indicators used by National Centres for monitoring coding.

- *Score agreement* refers to the proportion of scores (generally the first digit of assigned codes, denoting full, partial, and no credit) from one coder that exactly matched the scores of other coders on an identical set of multiple-coded responses for an item (including scores on partial credit item responses). Agreement can vary from 0 (0% agreement) to 1 (100% agreement). Each country/economy was expected to meet a scoring *standard* within-country and across-country proportion of at least 85% agreement on each item or coder in Mathematics, Reading, Science, and Financial Literacy; this standard was set to 70% agreement for Creative Thinking. Further, an average domain-level standard across all items in a domain of 92% was expected, except for Creative Thinking, which was also set to 70%. The design called for a minimum of one-hundred responses for each item to be multiple coded for the calculation of within-country score agreement; when fewer than 100 responses for an item in a particular country-by-language group were collected, as was the case of small samples, all responses were multiple coded. Additionally, ten (paper-based) or thirty (computer-based) English responses for each item were anchor coded for the calculation of across-country score agreement.
- *Coding category distribution* refers to the distributions of coding categories (such as “full credit”, “partial credit” and “no credit”) assigned by a coder to two sets of responses: a 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.

During coding, the formula used to by the OECS to calculate ongoing interrater agreement was:

$$R_{ji} = \frac{G_{ji} \left( \frac{N - A}{N} \right)}{D_{ji}(C - 1)} + \frac{A}{N} \quad \text{Formula 15.1}$$

where  $R_{ji}$  is the calculated agreement rate for coder  $C_j$  for item  $i$ ,  $N$  is the total number of responses for item  $i$ ,  $A$  is the number of automatically coded responses for item  $i$ ,  $C$  is number of coders for the item,  $G_{ji}$  is the number of agreed codes for coder  $C_j$  for item  $i$  (max =  $(C-1)$ ), and  $D_{ji}$  is the number of multiple-coded responses for item  $i$  coded by coder  $j$  so far (at the end of coding, this will equal 100 in a standard sample). The OERS reports calculated agreement similarly, with the exception that no responses were automatically coded (so,  $A = 0$ , simplifying the equation).

### **Score agreement across countries/economies, languages, and items**

Scoring reliability was again reviewed by the PISA contractor following the completion of coding to check for scoring consistency of human-coded CR items within and across countries participating in PISA 2022. For comparability among country-by-language groups and between multiple-coded student responses and anchor-coded English responses, the proportion of automatically coded responses were disregarded, and only the scoring reliability of human coders was considered. The reliability studies included 78 CBA countries/economies, resulting in 124 country-by-language groups. One PBA country, and three new PBA countries, each with one language group. In total there were 128 country-by-language groups across modes of assessment.

In a review of country-level data, quality and consistency of score agreement within and across country-by-language groups was evaluated. High score agreement is generally reflective of quality coding: that national and international coder trainings were well-implemented, coding guides were reflective of the student responses, such that scores could be consistently applied, and the scores applied on human-coded CR items are reliably accurate. All country-by-language groups were reviewed to see if the score agreement standard was met on all items and domains. Annex Table 15.A.4, Annex Table 15.A.5 and Annex Table 15.A.6 report the domain-level score agreement for all PISA 2022 participating countries and economies.

Overall, the majority of country-by-language groups administering the CBA met the domain-level within-country score agreement standard of 92% (or 70% in Creative Thinking):

- In Mathematics, 98.4% of country-by-language groups met the domain-level scoring standard; those below averaged 91.0% score agreement on multiple-coded responses.
- In Reading, 89.5% of country-by-language groups met the domain-level scoring standard; those below averaged 90.7% score agreement on multiple-coded responses.
- In Science, 73.4% of country-by-language groups met the domain-level scoring standard; those below averaged 90.7% score agreement on multiple-coded responses.
- In Financial Literacy, 86.7% of country-by-language groups met the domain-level scoring standard; those below averaged 91.4% score agreement on multiple-coded responses.
- In Creative Thinking, 97.0% of country-by-language groups met the domain-level scoring standard; those below averaged 69.6% score agreement on multiple-coded responses.

Note that in some cases, 100% score agreement was observed in certain country-by-language groups. This is more likely to occur when the number of responses being multiple coded is fewer than the recommended 100 student responses, usually due to a small sample size.

Quality in the coding of the English anchor responses is also important for ensuring that the coding guides have applied in the same way across countries/economies and language groups. Most country-by-language groups administering the CBA also met the relevant domain-level across-country score agreement standard of 85% (or 70% in Creative Thinking):

- In Mathematics, 92.7% of country-by-language groups met the domain-level scoring standard; those below averaged 86.0% score agreement on 30 anchor responses.
- In Reading, 87.1% of country-by-language groups met the domain-level scoring standard; those below averaged 86.0% score agreement on 30 anchor responses.
- In Science, 68.5% of country-by-language groups met the domain-level scoring standard; those below averaged 88.4% score agreement on 30 anchor responses.
- In Financial Literacy, 83.3% of country-by-language groups met the scoring standard; those below averaged 90.0% score agreement on 30 anchor responses.

- In Creative Thinking, 97.0% of country-by-language groups met the scoring standard; those below averaged 63.7% score agreement on 30 anchor responses.

Finally, all paper-based countries met the standard for across- and within-country score agreement all domains.

Annex Table 15.A.7 summarizes Annex Table 15.A.5, Annex Table 15.A.6 and Annex Table 15.A.6, providing an overall breakdown of score agreement of items by domain.

Across most domains and modes of assessment, across-country score agreement tended to be slightly lower than the within-country agreement by domain in the majority of country-by-language groups. This may be expected because, compared to multiple-coding, there are fewer bilingual coders (only two from the coding team) and fewer anchor-coded responses contributing to the calculation of agreement. However, the difference between multiple-coding and anchor-coding agreement by domain is generally minimal across country-by-language groups. In the domains of Mathematics, Reading, Science, and Financial Literacy, there was about 1-3% difference between the within-country agreement and the across-country agreement at the domain level in country-by-language groups, with only a few exceptions. In Creative Thinking, the domain level difference in agreement was closer to 7%, but with a lower threshold for standard of agreement, there is more room for fluctuation in agreement statistics, so this can also be expected.

### ***Coder-level score agreement***

Coder quality was also reviewed, particularly the percentage of coders in a country-by-language group that did not meet the standard level of agreement on across several items. Annex Table 15.A.8 and Annex Table 15.A.9 summarize overall coder quality and the impact of coder quality by item. In general, the coding standard indicates that all coders should agree with their fellow coders at least 85% of the time on each item, except in Creative Thinking, in which 70% score agreement was considered acceptable. Annex Table 15.A.8 shows the percentage of coders who were unable to reach the 85% agreement threshold on 20% or more items assigned to them. In Mathematics, 2.5% of coders agreed with their fellow coders less than 85% of the time on at least 20% of new item responses selected for multiple coding, and 0.8% were below this standard for trend item responses; this was also true of 8.8% of coders in Reading, 3.1% of coders in Science, and 0.7% of coders in Financial Literacy. In Creative Thinking, 15.7% of coders agreed with their fellow coders less than 70% of the time on at least 20% of responses.

Because coder quality is reflected at the item level, the percentage of items in the domain over which two or more coders did not meet the standard level agreement on that item was also evaluated, and the results are presented in Annex Table 15.A.9. Because there are a varying number of items in each domain, this table expresses the percentage of cases across all country-by-language groups. In other terms, however, about half of the CBA country-by-language groups had one new Mathematics item for which two coders did not meet the established 85% score agreement, and a fraction of that had this issue with a trend Mathematics item. Most country-by-language groups would have had about two reading items for which at least two coders did not meet the scoring standard, and in science, one item. About a third of all groups administering financial literacy would have had two coders below the standard on one item, and in Creative Thinking, all groups would have had about two items for which two or more coders did not reach 70% score agreement. There were no items across the paper-based domains for which there were two or more coders below the standard of score agreement on an item. These results overall suggest that any significant coding issues that may have arisen during coding were resolved at the National Centres.

### ***Item-level agreement***

The scales on which the PISA statistical framework is built are only as good as the scores used to establish them, so the overall agreement on student responses was also reviewed at the item-level, taking into

account the proportion of responses that could be automatically coded. Here, the interest is to determine the proportion of items in a country-by-language group that did not meet the standard level of agreement. Again, at the item level, the score agreement standard was set to 85% in all domains and modes of assessment, except for Creative Thinking, in which the standard was set to 70% score agreement. These standards were met for most items in each country-by-language group. Annex Table 15.A.10 shows the number of country-by-language groups that had either no items in a domain ( $n = 0$ ) below the standard, between one and five items ( $1 \leq n \leq 5$ ), or up to ten items ( $6 \leq n \leq 10$ ) in a domain below the score agreement standard. In the paper administration, all countries met the scoring standard on all items in all three domains. In the computer administration, all country-by-language groups met the standard on all items in Science and Financial Literacy. In Mathematics, one country-by-language group had items that failed to meet the standard, in Reading, four groups, and in Creative Thinking, five country-by-language groups had 1-5 items below the standard, and two had 6-10 items below the standard.

## Machine-supported coding system

During the 2022 cycle, the CBA coding teams were able to benefit from the use of a machine-supported coding system (MSCS). The MSCS operates effectively due to a high response regularity among collected student data. Consider that, although an item's response field is open-ended, there is a commonality among students' raw responses, meaning that the same or similar correct or incorrect responses can be expected regularly throughout coding (Yamamoto et al., 2017<sup>[2]</sup>; 2018<sup>[3]</sup>). High regularity in responses means that variability among all responses for an item is small, and a large proportion of identical responses can receive the same code when observed a second or third time. In such cases, human coding can be replaced by machine coding, greatly reducing the human coding burden and minimizing the error present in human-coded data, often associated with fatigue or carelessness.

Unlike commonly used automated scoring systems that generally involve algorithms, the MSCS relies entirely on text data that have already been human coded in past PISA cycles and during the field trial. These observed text responses and their associated verified codes from past administrations are stored in a *Coded Unique Response* (CUR) pool for each country-by-language group. In order for a text response to receive a verified code and be added to the CUR pool, the response must have appeared at least five times, and coders must have 100% agreement on the code to apply. The MSCS approach parallels automated scoring in the sense that a scoring model is first trained on existing historic data (2015 and 2018 PISA cycles and the 2022 field trial) and then applied to future data (2022 main survey). When raw student responses are received, and before they are distributed to human coders in the OECS, they are first checked to see if the MSCS can automatically apply a code. Raw responses fall into one of three categories: 1) nonresponse, 2) responses with verified coding in the CUR pool, and 3) infrequent or unseen responses that require human judgment. The MSCS can be applied to the first two categories. Human coding would only be required for unique, unseen responses (3). The MSCS is specific to each country-by-language group; responses that are identified for automatic coding are not shared among country-by-language groups. In brief, the MSCS identifies blank responses and the exact same responses that have been previously coded by humans and automatically applies the appropriate code, minimizing the need to score responses that have already been added to the database (Yamamoto et al., 2017<sup>[2]</sup>; 2018<sup>[3]</sup>; OECD, 2018<sup>[4]</sup>).

### **Reduction of human-coding burden as the result of the MSCS**

Annex Table 15.A.11 and Annex Table 15.A.12 summarize the efficiency of the MSCS with the reduction of human-coding burden in the PISA 2022 field trial and main survey. The tables summarize the percentage of responses coded by the MSCS and by human coders across all items in four domains (mathematics, reading, science, and financial literacy) and across country/economy language groups using mean and

median. Given that the distribution of proportions for each item per group can be skewed, medians are reported in addition to the mean values.

The first two columns under the “machine-coded” header, *CUR* and *nonresponse*, indicate the average and median percentage of responses across CBA items that were automatically coded by the MSCS as either a nonresponse or a verified response (correct – full and partial credit – and incorrect). The total of these values is also presented, which can be compared to the percentage of human-coded responses, noted in the first column. Note that without the MSCS, all of the responses to CR items would have had to be coded by humans, including nonresponses. On average, across items and country-by-language groups, the coding burden for human coders was reduced for the 2022 field trial from a low of approximately 14% on new Mathematics items to a high of 31% on trend Mathematics items. For the 2022 main survey, the coding burden was reduced by a low of approximately 7% in Creative Thinking, for which only nonresponses were coded by the MSCS, to a high of 35% (about 15% CUR and 20% nonresponse) on trend Mathematics items.

For both field trial and main survey, approximately 7% to 20% of the total responses (on average) across all domains were empty responses and were automatically coded by the system. On new items, where no historic data were available, the MSCS reduced coding burden for human coders by 12% to 14% in Mathematics and 7% to 15% in Creative Thinking. For new Mathematics items that received modification following the field trial, only empty responses were automatically coded during the main survey, which may explain why only 5% of new Mathematics responses were automatically coded through the CUR pool. Because of the format and generally graphical nature of the Creative Thinking domain, only empty responses were automatically coded by the MSCS in both the field trial and the main survey. Overall, a similar or slightly higher percentage of responses were coded in each of the core domains in the 2022 PISA cycle than in the previous cycle.

## References

- OECD (2018), *PISA 2018 Technical Report*, PISA, OECD Publishing, Paris, [4]  
<https://www.oecd.org/pisa/data/pisa2018technicalreport/>.
- Shin, H., M. von Davier and K. Yamamoto (2019), “Investigating Rater Effects in International Large-Scale Assessments”, in Veldkamp, B. and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement: Methodology of Educational Measurement and Assessment*, Springer, Cham. [1]
- Yamamoto, K. et al. (2018), “Development and implementation of a machine-supported coding system for constructed-response items in PISA”, *Psychological Test and Assessment Modeling*, Vol. 60/2, pp. 145-164. [3]
- Yamamoto, K. et al. (2017), “Developing a machine-supported coding system for constructed-response items in PISA”, *ETS Research Report*, No. RR-17-47, Educational Testing Service, Princeton, NJ, <https://doi.org/10.1002/ets2.121>. [2]

# Annex 15.A. Detailed Overview of the Coding Process

**Annex Table 15.A.1. Chapter 15: Comprehensive Analysis of Coding Practices**

Tables	Title
Table 15.A.2	Number of cognitive items by domain, item format, and coding method
Table 15.A.3	CBA coding number of coders by domain
Table 15.A.4	PBA and New PBA number of coders by domain
Web Table 15.A.5	Summary of within- and across-country (%) scoring agreement for CBA participants for reading, mathematics and science
Table 15.A.6	Summary of within- and across-country (%) agreement for Financial Literacy and Creative Thinking domains
Table 15.A.7	Summary of within- and across-country (%) scoring agreement for Paper-based countries
Table 15.A.8	Average item-level score agreement (across country-language groups) by domain
Table 15.A.9	Percentage of coders whose scoring was below the standard inter-rater agreement on 20% or more of items, averaged across countries
Table 15.A.10	Percentage of items in a domain with at least two coders below the standard scoring agreement on the item (in the same country-by-language group)
Table 15.A.11	Number of country-language groups with score agreement below the domain standard
Table 15.A.12	Percentage of responses coded by the MSCS and by human coders across countries in the 2022 field trial
Table 15.A.13	Percentage of responses coded by the MSCS and by human coders across countries in the 2022 main survey

StatLink  <https://stat.link/p4cuhz>

**Annex Table 15.A.2. Number of cognitive items by domain, item format, and coding method**

			Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
<b>CBA</b>	Human Coded	Constructed Response	19	16	64	32	16	34
	Computer Scored	Simple Multiple Choice	80	18	104	33	12	0
		Complex Multiple Choice	35	14	27	47	14	2
		Constructed Response	26	26	2	3	4	0
		<b>Total</b>	<b>160</b>	<b>74</b>	<b>197</b>	<b>115</b>	<b>46</b>	<b>36</b>
<b>PBA</b>	Human Coded	Constructed Response		38	51	32		
	Computer Scored	Simple Multiple Choice		18	27	29		
		Complex Multiple Choice		12	9	24		
		Constructed Response		3	0	0		
		<b>Total</b>		<b>71</b>	<b>87</b>	<b>85</b>		
<b>New PBA</b>	Human Coded	Constructed Response		40	37	9		
	Computer Scored	Simple Multiple Choice		16	24	34		
		Complex Multiple Choice		8	5	23		
		Constructed Response		0	0	0		
		<b>Total</b>		<b>64</b>	<b>66</b>	<b>66</b>		

**Annex Table 15.A.3. CBA coding number of coders by domain**

	Recommended Number of Coders by Number of Students Assessed				Example Workload*		
	< 4,500	4,501 – 8,000	8,001 – 13,000	> 13,000	Coders	Expected Coding Days	Responses per Coder
Mathematics	2 – 3	4 – 5	6 – 9	10 – 12	8	7.1	6,853
Reading	2 or 4	4 or 8	8 or 12	12 – 32	8	7.4	5,194
Science	2 – 3	4 – 5	6 – 9	10 – 12	8	6.3	6,107
Financial Literacy	2 – 3	4 – 5	6 – 9	10 – 12	8	4.8	4,691
Creative Thinking	2 – 3	4 – 5	6 – 9	10 – 24	8	8.9	5,974

Note: Example assumes a main sample size of 6 300 and a Financial Literacy sample size of 1 650.

Example assumes that coders in the core domains and Financial Literacy would be able to code approximately 1 000 responses per day, and coders in the Creative Thinking domain would be able to code approximately 700 responses per day.

**Annex Table 15.A.4. PBA and New PBA number of coders by domain**

		Example Workload		
		Coders	Expected Coding Days	Responses per Coder
PBA	Mathematics	6	14	10,418
	Reading	6	15	7,733
	Science	6	9	3,328
New PBA	Mathematics	6	17	11,667
	Reading	6	16	10,500
	Science	6	4	2,625

Note: Example assumes that coders would be able to code approximately 1 000 responses per day.

**Annex Table 15.A.5. Summary of within- and across-country (%) agreement for Financial Literacy and Creative Thinking domains**

	Country/Economy - Language	Within-country		Across-country	
		Financial Literacy	Creative Thinking	Financial Literacy	Creative Thinking
OECD	Australia - English		74.6%		88.7%
	Austria - German	93.0%		93.5%	
	Belgium - Dutch	96.6%	84.2%	96.6%	91.2%
	Belgium - French		76.8%		87.5%
	Belgium - German		76.8%		88.9%
	Canada - English	91.4%	76.2%	95.9%	91.0%
	Canada - French	92.6%	75.8%	94.6%	87.7%
	Chile - Spanish		76.7%		87.5%
	Colombia - Spanish		81.9%		86.7%
	Czech Republic - Czech	94.9%	89.7%	96.5%	91.7%
	Denmark - Danish	94.1%	86.8%	95.6%	90.6%
	Denmark - Faroese		98.1%		92.1%
	Estonia - Estonian		83.9%		96.9%
	Estonia - Russian		78.1%		86.6%
	Finland - Finnish		83.0%		92.9%
	Finland - Swedish		96.2%		93.9%
France - French		84.1%		90.7%	



		Germany - German		86.0%		88.7%
		Greece - Greek		80.6%		87.4%
		Hungary - Hungarian	92.5%	80.7%	95.5%	92.7%
		Iceland - Icelandic		80.3%		93.0%
		Israel - Arabic		83.2%		92.9%
		Israel - Hebrew		80.6%		93.0%
		Italy - German	95.6%	82.2%	93.4%	94.2%
		Italy - Italian	92.8%	91.1%	95.0%	96.4%
		Korea - Korean		85.7%		87.2%
		Latvia - Latvian		86.1%		85.8%
		Latvia - Russian		86.4%		88.9%
		Lithuania - Lithuanian		90.4%		95.6%
		Lithuania - Polish		89.0%		93.0%
		Lithuania - Russian		89.7%		95.0%
		Mexico - Spanish		80.7%		83.6%
		Netherlands - Dutch	92.0%	77.0%	92.2%	89.4%
		New Zealand - English		80.3%		91.0%
		Norway - Bokmål	95.4%		97.1%	
		Norway - Nynorsk	96.3%		97.3%	
		Poland - Polish	93.3%	79.5%	95.3%	91.0%
		Portugal - Portuguese	90.9%	81.4%	93.7%	85.4%
		Slovak Republic - Hungarian		98.1%		90.4%
		Slovak Republic - Slovak		87.3%		88.9%
		Slovenia - Slovenian		85.1%		89.2%
		Spain - Basque	95.0%	69.9%	87.4%	79.8%
		Spain - Catalan	92.6%	75.6%	93.3%	82.3%
		Spain - Galician	92.4%	72.5%	92.4%	81.0%
		Spain - Spanish	92.1%	69.0%	91.1%	83.6%
	*	Spain - Valencian	100.0%	82.8%	93.9%	83.6%
		United States - English	95.4%		97.7%	
Partners		Albania - Albanian		93.2%		83.0%
		Baku (Azerbaijan) - Azeri		76.6%		68.8%
		Baku (Azerbaijan) - Russian		77.3%		74.8%
		Brazil - Portuguese	99.3%	86.8%	98.8%	95.7%
		Brunei Darussalam - English		76.3%		87.8%
	Bulgaria - Bulgarian	91.3%	81.0%	96.4%	89.2%	
	Chinese Taipei - Chinese		79.1%		86.0%	
	Costa Rica - Spanish	93.1%	80.0%	94.3%	82.6%	
	Croatia - Croatian		94.8%		85.8%	
	Cyprus - English		87.6%		90.4%	
	Cyprus - Greek		78.0%		87.8%	
	Dominican Republic - Spanish		90.7%		62.7%	
	El Salvador - Spanish		84.8%		77.2%	
	Hong Kong (China) - Chinese		94.0%		95.1%	
	Hong Kong (China) - English		97.2%		94.1%	
	Indonesia - Indonesian		83.4%		84.7%	
	Jamaica - English		69.8%		86.0%	
	Jordan - Arabic		83.3%		83.4%	
	Kazakhstan - Kazakh		83.1%		89.2%	
	Kazakhstan - Russian		87.0%		89.7%	
	Macao (China) - Chinese		93.2%		93.6%	
	Macao (China) - English		91.8%		93.6%	
	*	Macao (China) - Portuguese		100.0%		84.9%

	Malaysia - English	94.4%	76.2%	97.8%	86.6%
	Malaysia - Malay	92.9%	81.0%	98.0%	86.8%
	Malta - English		77.0%		87.6%
	Malta - Maltese		81.4%		88.1%
	Mongolia - Mongolian		81.8%		86.2%
	Morocco - Arabic		81.7%		88.6%
	Morocco - French		82.9%		87.4%
	North Macedonia - Macedonian		86.3%		89.3%
	Palestinian Authority - Arabic		95.9%		86.2%
*	Palestinian Authority - English		98.5%		86.5%
*	Panama - English		95.2%		77.8%
	Panama - Spanish		97.1%		59.5%
	Peru - Spanish	96.2%	87.5%	96.1%	91.4%
	Philippines - English		87.8%		90.9%
	Qatar - Arabic		93.2%		86.9%
	Qatar - English		100.0%		87.4%
	Republic of Moldova - Romanian		93.3%		100.0%
	Republic of Moldova - Russian		92.7%		99.7%
	Romania - Hungarian		79.4%		85.8%
	Romania - Romanian		81.0%		88.2%
	Saudi Arabia - Arabic	91.8%	83.7%	90.8%	88.5%
	Saudi Arabia - English	92.0%	97.2%	92.3%	90.0%
*	Serbia - Hungarian		89.7%		90.7%
	Serbia - Serbian		85.7%		90.8%
	Singapore - English		86.7%		91.3%
	Thailand - Thai		92.0%		93.1%
*	Ukraine - Russian		100.0%		87.7%
	Ukraine - Ukrainian		82.0%		90.0%
	United Arab Emirates - Arabic	93.1%	79.1%	90.3%	79.5%
	United Arab Emirates - English	92.8%	78.3%	90.3%	81.0%
	Uruguay - Spanish		80.6%		92.5%
*	Uzbekistan - Karakalpak		88.2%		84.7%
	Uzbekistan - Russian		91.8%		91.1%
	Uzbekistan - Uzbek		88.6%		87.0%

\* Denotes a country-language group which assessed fewer than 200 students; therefore, there are fewer multiple coded responses contributing to the calculation of agreement in these groups.

Note: Originally assigned codes for Creative Thinking were rescored for some items during scaling; agreement in this table reflects the original human scoring.

**Annex Table 15.A.6. Summary of within- and across-country (%) scoring agreement for Paper-based countries**

	Country/Economy - Language	Within-country Agreement			Across-country Agreement		
		Mathematics	Reading	Science	Mathematics	Reading	Science
<b>Partners</b>	Guatemala - Spanish	99.9%	99.8%	99.1%	98.5%	97.5%	96.5%
	Cambodia - Khmer	99.5%	99.5%	99.1%	99.6%	99.2%	99.3%
	Paraguay - Spanish	99.3%	97.5%	97.4%	99.0%	97.3%	97.2%
	Viet Nam - Vietnamese	99.9%	99.3%	99.7%	98.2%	94.0%	96.6%

**Annex Table 15.A.7. Average item-level score agreement (across country-language groups) by domain**

		Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
CBA	Multiple-coded	95.4%	97.5%	95.4%	94.0%	93.9%	85.0%
	Anchor	93.8%	97.6%	94.8%	92.7%	94.4%	87.9%
PBA and New PBA	Multiple-coded		99.7%	99.0%	98.8%		
	Anchor		98.8%	97.0%	97.4%		

**Annex Table 15.A.8. Percentage of coders whose scoring was below the standard inter-rater agreement on 20% or more of items, averaged across countries**

	Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
CBA	2.5%	0.8%	8.8%	3.1%	0.7%	15.7%
PBA and New PBA		0.0%	0.0%	0.0%		

Note The standard is set to 85% agreement in mathematics, science, reading, and financial literacy; in Creative Thinking, it is set to 70% agreement.

**Annex Table 15.A.9. Percentage of items in a domain with at least two coders below the standard scoring agreement on the item (in the same country-by-language group)**

	Mathematics (New)	Mathematics (Trend)	Reading	Science	Financial Literacy	Creative Thinking
CBA	3.0%	1.1%	4.2%	4.2%	2.1%	7.5%
PBA and N-PBA		0.0%	0.0%	0.0%		

**Annex Table 15.A.10. Number of country-language groups with score agreement below the domain standard**

	N Items below the Standard	Mathematics (New)	Mathematics (Trend)	Science	Reading	Financial Literacy	Creative Thinking
CBA	$N = 0$	123	123	124	120	30	93
	$1 \leq N \leq 5$	1	1	0	2	0	5
	$6 \leq N \leq 10$	0	0	0	2	0	2
PBA and New PBA	$N = 0$		4	4	4		
	$1 \leq N \leq 5$		0	0	0		
	$6 \leq N \leq 10$		0	0	0		

Note: The standard is set to 85% agreement in Mathematics, Science, Reading, and Financial Literacy and 70% in Creative Thinking.

**Annex Table 15.A.11. Percentage of responses coded by the MSCS and by human coders across countries in the 2022 field trial**

		Human Coded		Machine Coded	
				CUR	Nonresponse
Mathematics (New)	Mean	85.81%	NA	14.19%	14.19%
	Median	93.85%	NA	6.15%	6.15%
Mathematics (Trend)	Mean	69.04%	11.94%	19.02%	30.96%
	Median	73.86%	1.59%	16.11%	26.14%
Reading	Mean	71.06%	10.70%	18.23%	28.94%

	Median	77.78%	0.00%	13.51%	22.22%
Science	Mean	78.44%	8.65%	12.90%	21.56%
	Median	81.48%	0.00%	8.44%	18.52%
Financial Literacy	Mean	84.42%	0.98%	14.60%	15.58%
	Median	87.09%	0.00%	12.27%	12.91%
Creative Thinking	Mean	85.23%	NA	14.77%	14.77%
	Median	89.91%	NA	10.09%	10.09%

Note: Mean values are the mean of the total percentage of responses coded by the MSCS within each domain across countries; median values are the median of those percentages across countries and, therefore, may not add up to 100%.

Note: CUR pool responses were not available for new items in Mathematics, Financial Literacy, or Creative Thinking.

### Annex Table 15.A.12. Percentage of responses coded by the MSCS and by human coders across countries in the 2022 main survey

		Human Coded	Machine Coded		
			CUR	Nonresponse	Total
Mathematics (New)	Mean	82.85%	4.76%	12.39%	17.15%
	Median	91.30%	0.00%	6.70%	8.70%
Mathematics (Trend)	Mean	65.22%	14.88%	19.90%	34.78%
	Median	69.53%	4.48%	16.67%	30.47%
Reading	Mean	71.32%	10.68%	18.00%	28.68%
	Median	78.20%	0.43%	13.67%	21.80%
Science	Mean	75.49%	11.42%	13.09%	24.51%
	Median	77.89%	1.70%	8.78%	22.11%
Financial Literacy	Mean	84.77%	2.99%	12.24%	15.23%
	Median	86.19%	0.00%	10.44%	13.81%
Creative Thinking	Mean	92.36%	NA	7.64%	7.64%
	Median	94.37%	NA	5.63%	5.63%

Note: Mean values are the mean of the total percentage of responses coded by the MSCS within each domain across countries; median values are the median of those percentages across countries and, therefore, may not add up to 100%.

Note: CUR pool responses were not available for some new items in Mathematics that had changes following the field trial; CUR pool responses were not applied in Creative Thinking.

# 16 Data Adjudication

## Introduction

Data adjudication is the process through which each national dataset is reviewed and a judgement about the appropriateness of the data for the main reporting goals is formed. The PISA Technical Standards (see Annex I) specify the way in which PISA must be implemented in each participating jurisdiction and adjudicated region. 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 PISA 2022 data for each of the adjudicated entities (i.e. the participating countries and economies – hereafter, “jurisdictions” – and the adjudicated regions) 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, including concerns about the quality of the data that were identified during scaling and in preparation for reporting
- the outcomes of the international coding review.

Not all regions (i.e. subnational jurisdictions that report their results separately) 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.

### ***PISA 2022 Technical Standards***

The areas covered in the PISA 2022 Technical Standards include several aspects connected to the implementation of PISA, including the definition and sampling of its target population on appropriate languages for testing, translation and adaptation of materials, school and student participation in the Field Trial and Main Survey, test administrations and handling of test materials, coding of responses, data management, privacy, and submissions, to cite a few key aspects. A comprehensive list of Technical Standards used for adjudication is available in Annex Table 16.A.1.

### ***Implementing the standards – quality assurance***

National Project Managers of participating jurisdictions 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- assurance procedures 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 2022 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 the data adjudication, information collected during both the quality-assurance and quality-monitoring activities was used to determine the level of compliance with the standards.

### ***Information available for adjudication***

The international contractors' quality monitoring of a participating jurisdiction's data collection is carried out from a range of perspectives during many stages of the PISA cycle. These perspectives include monitoring a participating jurisdiction's adherence to the deadlines, communication from the sampling contractor about each participating jurisdiction's sampling plan, information from the linguistic 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 said non-compliance 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 Technical 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 areas:

- international contractors' administration and management: information relating to administration processes, agreement of adaptation spreadsheets, submission of information.
- translation: information from linguistic verification of test items, questionnaire items, and the test administration script.
- sampling: information from the submitted data such as school and student response rates, exclusion rates and eligibility problems.
- school-level materials: information from the agreement of adaptations to test administration procedures and field operations.
- student materials: information from the pre- and post- Main Survey final optical checks of MS test booklets and background questionnaires.
- National Centre operations: School Coordinator, Test Administrator or School Associate trainings; information gathered through interviews conducted during meetings of National Project Managers or at other times.
- PISA Quality Monitors (PQMs): co-ordination of PISA Quality Monitor activities including recruitment; information gathered via the Data Collection Forms from PQMs and through their interactions with School Co-ordinators and Test Administrators.
- data cleaners: issues identified during the data cleaning checks and from data cleaners' reports.

- data processing: issues relating to the eligibility of students tested; issues identified in the coder query service and training of coders.
- data analysis: information from item level reports, from the Field Trial data, and from data cleaning steps, including consistency checks.
- questionnaire data: issues relating to the questionnaire data in the national questionnaire reports provided by the international contractor.
- Main Survey and Field Trial Reviews: information provided by the National Project Managers in the Field Trial and Main Survey Review 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 international contractor in charge of coordinating PISA implementation (Core A, see Chapter 1) where it was aggregated by the international project manager at the contractor. The PQM reports contained data related to test administration in selected schools, and the PISA quality monitoring data were collected independently of the National Project Manager.

### **Data adjudication process**

Data adjudication is the process through which each national dataset is reviewed and a judgement about the appropriateness of the data for the main reporting goals is formed. The different steps in the data adjudication process ensure that the final judgement is transparent, based on evidence, and defensible.

The data adjudication process achieved this through the following steps:

- Step 1: International contractors collected quality-assurance and quality monitoring data throughout the survey administration period. The international project manager compiled this information into an adjudication database that was updated or amended as new information arose and provided an overview of the national implementation of PISA throughout the cycle.
- Step 2: The international project manager compiled individual reports for each jurisdiction that contained quality-assurance data for key areas of project implementation.
- Step 3: The international project director, together with the international contractor leads, identified data issues that were in need of adjudication. Where necessary, the relevant National Project Manager was contacted to provide additional information. After this stage, for each dataset, a summary report detailing whether and how the PISA Technical Standards had been met was drafted.
- Step 4: The PISA Adjudication Group, formed by representatives of the OECD, of international contractors, the Technical Advisory Group and the Sampling Referee, reviewed the summary reports to recommend adequate treatment of the data from each adjudicated entity in international PISA products (database and reports).
- Step 5: The recommendations of the PISA Adjudication Group were presented to the PISA Governing Board representatives and to the countries concerned.

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 I and Annex Table 16.A.1).

Information was collected in relation to these standards through a variety of instruments and information sources:

- through PISA Quality Monitor reports
- through the Field Trial and Main Survey reviews submitted by National Centres
- through information negotiated and stored on the communications portal for PISA 2022
- 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
- through the data cleaning and data submission process.

For PISA 2022, an adjudication database was developed to capture, summarise, and store the most important information derived from these various sources. International contractor staff 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 appropriate and comparable 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, 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. 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.

The adjudication process for PISA 2022 had an increased challenge imposed by the COVID-19 global pandemic that caused school closures worldwide. The onset of the pandemic was in early 2020 – right as early testing countries were about to implement their Field Trial data collections for the – then named – PISA 2021 cycle.

School closures affected education systems differently, but overall, significantly disrupted survey operations, which led to a decision by the PISA Governing Board to postpone data collection for one year, thus renaming the cycle PISA 2022. As schools reopened and instruction was normalized throughout 2021 and 2022, data collection for PISA resumed, but nonetheless required a degree of reactivity and flexibility from education systems and international contractors. Indeed, as schools reopened at an uneven pace throughout jurisdictions and attempted to get back on track for the rest of the (sometimes expedited) school year, participation of said schools in PISA 2022 Main Survey data collection could not be taken for granted, neither could access from Test Administrators to students, or even high attendance of the latter.

These limitations compelled some changes to Standard 1.3 (assessment period), namely:

- Extension of the assessment period beyond 56 days where students remain within the PISA-eligible age range would be agreed to with the OECD's implicit approval.
- Extension of the assessment period that would not exceed the allowed 56 days but would result in assessed students who are outside of the PISA-eligible age range **by less than a week** would be agreed to with the OECD's implicit approval.



- Extension of the assessment period that would both exceed 56 days AND result in assessed students who are outside of the PISA-eligible age range will require further consultation with the contractors and the OECD before approval of such a deviation would be granted.

The changes were proposed by the international contractors, endorsed by the PISA Technical Advisory Group (TAG) on its December 2021 meeting, and implemented throughout PISA 2022 MS data collection. All participating jurisdictions managed to successfully conclude MS data collection using this added flexibility.

### **Data adjudication outcomes**

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 the data be declared fit for use with explicit cautions advised regarding its representativity of the jurisdiction's student cohort, or international comparability of its results
- that some data be removed for a particular participating jurisdiction 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 rescoring 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 2022 database.

Throughout PISA 2022, 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**

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 participating jurisdiction was adjudicated at the adjudication group meeting in July 2023, the quality assurance and monitoring processes outlined earlier in this chapter, foreseen in the Technical Standards, and described throughout this report had been implemented. Data adjudication focused on residual issues that remained after these quality assurance processes had been carried out.

Overall, the Adjudication Group's review suggests good adherence of national implementations of PISA to the technical standards in spite of the challenging circumstances that affected not only PISA operations but schooling more generally during the COVID-19 pandemic. Thanks to the reactivity and flexibility of participating countries and international contractors, to carefully constructed instruments, to a test design that is aligned to the main reporting goals and is supported by adequate sample design, and to the use of appropriate statistical methods for scaling, population estimates are highly reliable and comparable across countries and time, and particularly with 2018 results.

Nevertheless, a number of deviations from standards were noted and their consequences for data quality were reviewed in depth. The following overall patterns of deviations from standards were identified:

- About one in five of all adjudicated entities had exclusion rates exceeding the limits set by the technical standards (Standard 1.7).

- Seven entities failed to meet the required school response rates, with three of them failing to meet the stricter level of 65% before replacement (Standard 1.11). This is in line with earlier cycles of PISA.
- There was a significant increase in the number of entities that failed to meet the required student response rates (Standard 1.12): 10 entities did not meet this standard.
- There were delays in data submission in a significant number of entities (Standard 19.1): 14 entities did not meet this standard, and 13 only partially met it. The Adjudication Group noted that delayed submissions may affect the quality of the international contractors' work; and if shorter reporting timelines are expected, it may no longer be possible to accommodate such delays.
- A large number of entities did not conduct the field trial as intended (Standard 3.1) or did not attend all meetings (Standard 23.1). While this may also be a consequence of the pandemic, the Adjudication Group noted that these violations may be particularly consequential for new participants and for less-experienced teams. The Group underlined the importance of attendance at coder training sessions for ensuring comparability of the data.

At the international level, these frequent deviations should guide future efforts of the PISA Governing Board, the OECD Secretariat and Contractors to review the corresponding standards, prevent future deviations from standards, or mitigate the consequences of such violations.

At the level of individual adjudicated countries, economies and regions, in most cases, these issues did not result in major threats to the validity of reports, and the data could be declared fit for use. Where school or student participation rates fell short of the standard and created a potential threat for non-response/non-participation bias, countries/economies were requested to submit non-response-bias analyses. The evidence produced by countries/economies (and in some cases, by the sampling contractor) was reviewed by the Adjudication Group.

The Adjudication Group reviewed and discussed major adjudication issues in June 2023. The major adjudication issues reviewed by the group fall under three broad categories: (1) exclusions and response rates, (2) invariance of item parameters, and (3) issues originating from the Chromebook online administration pilot.

## **Overview of exclusion and response rate issues**

### *Exclusion rates*

PISA Technical Standard 1.7 states that the target population - the population that sits PISA - covers at least 95% of the desired population, the one for which broader conclusions from the assessment are sought. This means that overall exclusions, at both school and student levels, cannot exceed 5%. Sixteen jurisdictions excluded students in excess of this threshold at varying degrees.

Data collection was severely disrupted by the onset of Russia's war of aggression on February 2022, meaning that only data for 18 out of Ukraine's 27 regions could be collected. Exclusions were computed with respect to the original sampling frame, covering the entire country. After February 2022, however, survey operations could not be completed successfully in the regions most affected by wartime disruptions adding to an exclusion rate of 36.1%. Results for the remaining regions were deemed fit for reporting, but comparisons with previous results should be made only with great caution, and with due consideration to the differences in target populations. The Adjudication Group recommended that the results are presented in such a way to alert readers to the difference in the target population between prior cycles and PISA 2022.

Exclusions in Denmark increased by a large margin, presenting a marked increase compared to previous cycles, at 11.6% in PISA 2022. The Adjudication Group noted that high levels of student exclusions may

bias performance results upwards. In Denmark, a major cause behind the rise appears to be the increased share of students with diagnosed dyslexia, and the fact that more of these students are using electronic assistive devices to help them read on the screen, including during exams. The lack of such accommodation in PISA led schools to exclude many of these students, meaning that rates are likely to fall should said accommodations, especially those supporting dyslexic students, are allowed.

Rather elevated rates were also observed in the Netherlands and in Latvia, at 8.4% and 7.9% of schools respectively. On the former participant, the Dutch National Centre submitted non-response bias analysis was submitted, analysing differences in performance and in other characteristics between responding schools and the total population of schools, as well as differences between replacement schools and originally sampled, but non-responding schools. This supported the case that no large bias would result from non-response; furthermore, given the available evidence, there is no clear indication about the direction of any residual bias.

The Adjudication Group noted that exclusions exceeded the acceptable rate by a small margin in:

- Croatia (5.4%), Lithuania (6.5%), and in the United States (6.1%), which showed a marked increase in exclusions due to students with functional or intellectual disabilities, which are also bound to fall in the presence of increased accessibility in future cycles. The Adjudication Group invited the national centres to investigate the reasons for this increase in exclusion rates and take remedial action for future cycles. It is expected that exclusion rates will fall again in the future, as a result.
- Australia (6.9%), Canada (5.8%), Estonia (5.9%), New Zealand (5.8%), Norway (7.3%), Scotland (6.6%), Switzerland (5.8%), and Türkiye (5.6%) where the exclusion rates observed in 2022 remained relatively close to exclusion rates observed in 2018.
- Sweden, (7.4%), which showed a marked decrease from the high levels observed in 2018. The Adjudication Group noted that this might be the combined result of falling rates of refugee students and of the national centre's further effort to ensure uniform application of guidelines across schools.

### *School response rates*

The PISA school response rate requirements are foreseen by Technical Standard 1.11, stating that all jurisdictions are to reach a weighted school response rate of 85%, which can be accomplished by administering PISA in replacement schools as needed. A comprehensive account of both jurisdictions and adjudicated regions' response rates either before and after replacement can be found in Annex Tables 13.A.3, 13.A.4, 13.A.5, and 13.A.6 respectively in Chapter 13 of this report.

Seven jurisdictions did not meet the 85% school participation threshold before replacements, with only two meeting the standard with the use of replacement schools. Nonetheless, the increase in participation of replacement schools was rather heterogeneous across jurisdictions. This lack of response, in particular at the school level, has the potential to induce bias in observed results, and thus further investigations in the form of a non-response bias analysis (NRBA) were conducted by affected jurisdictions supported by the international sampling contractor.

Such is the case of the Netherlands, where 66% of sampled schools responded, a share that increased to 90% upon replacement. A NRBA was submitted, analysing differences in performance and in other characteristics between responding schools and the total population of schools, as well as differences between replacement schools and originally sampled, but non-responding schools. This supported the case that no large bias would result from non-response; furthermore, given the available evidence, there is no clear indication about the direction of any residual bias.

Similarly, in Chinese Taipei, where the standard was nearly met before and after replacement (83% before replacement, 84% after), a thorough NRBA was produced, using school-level achievement data as

auxiliary information, which provided convincing evidence that the potential bias is minimal after non-response adjustments are considered.

The effect of replacement schools was less pronounced in the United States, where 51% of schools responded before replacement, with the school response rate going up to 62% once replacement schools were invited to participate. Participation rates thus missed the standard by a significant margin, with particularly low participation rates among private schools (representing about 7% of the student population). A NRBA was submitted, indicating that, after replacement schools and non-response adjustments are considered, a number of characteristics (not including direct measures of school performance) are balanced across respondents and non-respondents. Based on the available information, it was not possible for the Adjudication Group to exclude the possibility of bias, nor to determine its most likely direction.

Four other jurisdictions: Canada (81% before replacement, 86% after), Hong Kong (China; 60% before replacement, 80% after), New Zealand (61% before replacement, 72% after), and the United Kingdom (excl. Scotland; 66% before replacement, 80% after) have also submitted NRBA's supporting the case that any bias resulting from school non-participation is most likely be negligible, and are discussed in further detail below, as student participation was also a concern.

### *Student response rates*

Technical Standard 1.12 states that student response rates must be in excess of 80% across responding schools. Students are not replaced in PISA, and thus only those sampled and present at the testing sessions are to sit the test. Student response rates for all jurisdictions and adjudication entities can be found at Tables 13.7 and 13.8 in Chapter 13 of this report.

Ten jurisdictions did not meet this standard. Albeit some observed rates were close to the 80% threshold, a downwards trend in student participation is of particular concern. Checking for this potential bias is particularly significant, as students' absence from school for the PISA 2022 cycle comes in the aftermath of a global pandemic with severe economic consequences, which might affect some students more than others.

In Malta the student response rate observed in PISA 2022 (79%), fell short of the standard by a small margin, but decreased significantly (from 86%) compared to PISA 2018. A thorough non-response bias analysis (NRBA) was produced, using student-level academic track variable as auxiliary information, along with demographic characteristics. Because students were tracked in the previous academic years based on their grades in mathematics, track information can be expected to correlate strongly with performance on the PISA test. The NRBA provides convincing evidence that the potential bias is minimal after non-response adjustments are taken into account.

Similarly, in the United Kingdom (excl. Scotland), student response rates decreased to 75% from 83% with respect to PISA 2018. School response rates also fell short of the target as discussed above. An informative NRBA was submitted, using external achievement data at student level as auxiliary information, along with demographic characteristics; the analysis was limited to England as the largest subnational entity within the UK, and thus covered over 90% of the intended sample. The analysis provided evidence to suggest a small residual upwards bias, after non-response adjustments are considered, driven entirely by student non-response while school non-participation did not result in significant bias. On the PISA scale, considering that the standard deviation in Scotland (in 2018) was about 95 score points in reading and mathematics, this could translate in a bias of approximately 9 or 10 points.

On the other hand, in Scotland, student response rates missed the standard by one percentage point but were otherwise similar to response rates in PISA 2018 (81%). A thorough non-response bias analysis was submitted, using several external achievement variables at student level as auxiliary information, along with demographic characteristics. The analysis provided evidence to suggest a residual upwards bias of

about 0.1 standard deviations, after non-response adjustments are taken into account. On the PISA scale, considering that the standard deviation in Scotland (in 2018) was about 95 score points in reading and mathematics, this could translate in an estimated bias of approximately 9 or 10 points. Given the similarity of response rates between 2018 and 2022, it cannot be excluded that a similar bias might be present in 2018 as well, and in many PISA 2022 participants whose response rates were similarly close to the target. For this reason, data were deemed to be comparable to previous cycles.

Decreases in student participation were more severe for other jurisdictions. In the challenging circumstances surrounding schooling in Panama in 2022 (teacher strikes, road blockades, and student absenteeism), student response rates decreased from 90% in PISA 2018 to 77% in PISA 2022. However, no NRBA was submitted; the PISA national centre explained that non-response was potentially related to the agitated school climate the students found themselves when returning to their schools after the strikes. A limited NRBA was prepared by the international sampling contractor, to compare respondent characteristics (both before and after nonresponse adjustment) to characteristics of the full eligible sample of students. This analysis suggested that (before non-response adjustments were taken into account), non-response was related to students' grade level, and to special needs status. Based on the available information, it is not possible to exclude the possibility of bias; considering the analyses on student non-response conducted in other countries, the residual bias after non-response adjustments are taken into account is likely to correspond to an upward bias.

In Canada, response rates decreased to 77% in PISA 2022 from 84% observed in PISA 2018. A thorough NRBA was submitted, with analyses conducted separately for each Canadian province, using students' academic achievement data as auxiliary information. School response rates also fell short of the target, driven by low participation rates in two provinces (Québec and Alberta). For these provinces, non-response bias was also examined at the school level. The analyses clearly indicate that school nonresponse has not led to any appreciable bias, but student nonresponse has given rise to a small upwards bias.

A similar decrease in participation was observed in Ireland, where student response rates decreased to 77% in PISA 2022 from 86% with respect to PISA 2018. A thorough NRBA was submitted, using external achievement data at student level as auxiliary information. The analysis provided evidence to suggest a residual upwards bias of about 0.1 standard deviations, after non-response adjustments are taken into account. On the PISA scale, considering that the standard deviation in Ireland ranged (in 2018) from 78 score points in mathematics to 91 score points in reading, this could translate in an estimated bias of approximately 8 or 9 points. The Adjudication Group also noted that the bias associated with trend and cross-country comparisons might be smaller, if past data or data for other countries are biased in the same direction.

Australia also observed a decline in student response rates, from 85% in PISA 2018 to 76% in PISA 2022. A technically sound NRBA was submitted; however, the strength of the evidence was limited by the fact that no external student-level achievement variables could be used in the analysis. Based on the available evidence, and on the experience of other countries participating in PISA, the Adjudication Group considered that while non-response adjustments likely limited the severity of non-response biases, a small residual upward bias could not be excluded.

Hong Kong (China) had a similar decrease from 85% in PISA 2018 to 75% in PISA 2022. School response rates also fell short of the standard (as they did in 2018). At the school level, the fact that a raw, but direct measure of school performance is used to assign schools to sampling strata (and therefore, differential non-response across strata is unlikely to cause bias), limits the risk of bias due to non-response. A NRBA was submitted; however, the strength of the evidence was limited by the fact that no external student-level achievement variables could be used in the analysis (only student grade information, already used in non-response adjustments, was available). The proxies for school and student achievement (school size and student grade) that were used in the analyses showed no or very limited relationship with participation rates. Nevertheless, based on the available evidence, and on the experience of other countries

participating in PISA, the Adjudication Group considered that while non-response adjustments likely limited the severity of non-response biases, a small residual upward bias could not be excluded.

New Zealand also experienced a decline in student participation and did not meet this standard in PISA 2022. Indeed, student response rates decreased from 83% in PISA 2018 to 72% in PISA 2022. School response rates also fell short of the target as shown above. A thorough and detailed NRBA was submitted, using external achievement data at student level, but also information on chronic absenteeism along with demographic characteristics to further support comparisons. The analysis provided evidence to suggest a residual upwards bias of about 0.1 standard deviations, after non-response adjustments are considered, driven entirely by student non-response with no discernible bias due to school non-response. The analysis also suggested that chronically absent students are over-represented among non-respondents in PISA. On the PISA scale, considering that the standard deviation in New Zealand ranged (in 2018) from 93 score points in mathematics to 106 score points in reading, this could translate in an estimated bias of approximately 10 points. The Adjudication Group also noted that the bias associated with trend and cross-country comparisons might be smaller, if past data or data for other countries are biased in the same direction.

A new participant in PISA 2022, Jamaica also observed student participant rates well below the standard, at 66%. A simple NRBA was submitted by the National Centre, analysing student response rates by school characteristics: this showed in particular lower response rates in schools located in rural areas. A limited NRBA was also prepared by the international sampling contractor, to compare respondent characteristics (both before and after nonresponse adjustment) to characteristics of the full eligible sample of students. This suggested that non-response was also related to students' grade level and gender (both variables are used in non-response adjustments). Based on the available information, it is not possible to exclude the possibility of bias; considering the analyses on student non-response conducted in other countries, the residual bias after non-response adjustments are taken into account is likely to correspond to an upward bias.

The Adjudication Group noted that a number of issues encountered during the Main Survey data collection could have been prevented, had Jamaica been able to do a full Field Trial, which was not possible due to COVID-related disruptions to schooling in 2021. In particular, enrolment information available to the national centre for school-level sampling often turned out to be imprecise; and low student participation rates could have been anticipated, had a regular Field Trial been conducted. As a result of inaccurate sampling frames and low student response rates, the achieved sample size for the Main Survey was well below target, and sampling errors for Jamaica are larger than desired. In spite of the violations of sampling standards, the Adjudication Group considered the data of sufficient quality for reporting if reports are annotated with appropriate notes of caution.

### ***Invariance of item parameters***

Albeit not a formal PISA Technical Standard for the 2022 cycle, the share of non-invariant items (i.e. presenting differential item functioning) was considered for adjudication, as measurement invariance underpins the international comparability of results, and thus, one of the main goals of PISA itself. During its 2021 December meeting, the PISA Technical Advisory Group (TAG) provided guidance on this matter<sup>1</sup> and set the share of two thirds of invariant items for each significant language group within a jurisdiction as a threshold for adjudication on whether there is sufficient alignment with the international PISA scale for results to be deemed comparable. Viet Nam did not meet this psychometric threshold.

In Viet Nam, mathematics and science scores were considered fit for reporting, given that for the vast majority of items, student responses were in line with the expectations derived from the experience of other countries participating in PISA. In reading however, a strong linkage to the international PISA scale could not be established as 40% of items in reading (35 of 87) were assigned unique (group-specific) parameters. The Adjudication Group noted that this lack of fit in reading might also reflect differences in construct

coverage of the PISA paper-based instrument used in Viet Nam and noted that this instrument will no longer be available in PISA 2025 and further cycles. Furthermore, in addition to item invariance, the Adjudication Group also noted that the response patterns in all subjects deviated significantly from those observed in Viet Nam in earlier cycles; for this reason, the Adjudication Group recommended breaking the trend for Viet Nam, and avoiding comparisons of scale scores to those reported in past cycles.

### **Chromebook pilot administration issues**

In PISA 2022, Iceland, and Norway participated in a pilot administration of online data collection using Chromebooks. Schools in both countries, especially those tested at the first or last days of testing windows in both countries, were affected by server outages during testing. This outage caused slowness and unresponsiveness for some students taking the cognitive assessment, thus resulting in inferior test conditions.

While the PISA Consortium solved this problem during the testing period, 579 students in Iceland (17.2% of the final student sample, unweighted) and 584 students in Norway (8.8%) were assessed on Chromebooks before the problem was solved. According to Iceland, test administrators reported the issue having affected at most 13% of the unweighted final sample (438 students). Data analyses for both countries indicated noticeable differences in the overall response time and overall response rate between students that took the test on days affected by technical problems and those tested on other days, but no noticeable differences were observed in the fit statistics, proficiency estimates or performance between the two testing sessions (first and second hour). In December 2022, the TAG reviewed these results and supported the proposal to keep the entire data for both countries, including that of days when Chromebook issues were reported, also considering that in PISA students are not penalised for having non-reached items at the end of each session. The Adjudication Group, in June 2023, confirmed that overall, the data, including those of students who took the test in these circumstances, were considered to be fit for reporting. The group noted that while it is not possible to exclude that the issue affected students' engagement and motivation to give their best effort, and therefore may have resulted in a small negative impact, their responses did show good fit with the model, and were not remarkably different from the performance of students in other schools.

## Notes

- 
1. Namely, the summary record of the PISA TAG meeting reads as follows: *For each major language of assessment within a participating country/economy, over two-thirds of items per domain are expected to be invariant from the international item parameters for the Field Test and the Main Survey. Cases with less than two-thirds of common items will undergo further review of their items and response data.*

## Annex 16.A. Data adjudication additional items

Annex Table 16.A.1. PISA 2022 Technical Standards considered in data adjudication

Area	Standard
Target population and sampling	<b>Standard 1.2:</b> Unless otherwise agreed upon only PISA-Eligible students participate in the test.
	<b>Standard 1.3:</b> Unless otherwise agreed upon, the testing period: <ul style="list-style-type: none"> <li>• is no longer than eight consecutive weeks in duration for computer-based testing participants,</li> <li>• is no longer than six consecutive weeks in duration for paper-based testing participants,</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 testing period in the previous PISA cycle</li> </ul> NOTE: TAG approved deviations to the testing period when necessary due to covid-19. a) Extension of assessment period beyond 8 weeks, students still in PISA-eligible age range b) Extension of assessment period does not exceed 8 weeks but students assessed may be outside of PISA-eligible age range by less than one week c) OECD and contractors pre-approved extension beyond 8 weeks that resulted in students outside the PISA-eligible age range being assessed.
	<b>Standard 1.7:</b> The PISA Defined Target Population covers 95% or more of the PISA Desired Target Population. That is, school-level exclusions and within-school exclusions combined do not exceed 5%.
	<b>Standard 1.8:</b> The student sample size for the computer-based mode is a minimum of 6300 assessed students, and 2100 for additional adjudicated entities, or the entire PISA Defined Target Population where the PISA Defined Target Population is below 6300 and 2100 respectively. The student sample size of assessed students for the paper-based mode is a minimum of 5250.
	<b>Standard 1.9:</b> The school sample size needs to result in a minimum of 150 participating schools, and 50 participating schools for additional adjudicated entities, or all schools that have students in the PISA Defined Target Population where the number of schools with students in the PISA Defined Target Population is below 150 and 50 respectively. Countries not having at least 150 schools, but which have more students than the required minimum student sample size, can be permitted, if agreed upon, to take a smaller sample of schools while still ensuring enough sampled PISA students overall.
	<b>Standard 1.10:</b> The minimum acceptable sample size in each school is 25 students per school (all students in the case of school with fewer than 25 eligible students enrolled).
	<b>Standard 1.11:</b> The final weighted school response rate is at least 85% of sampled eligible and non-excluded schools. If a response rate is below 85% then an acceptable response rate can still be achieved through agreed upon use of replacement schools.
	<b>Standard 1.12:</b> The final weighted student response rate is at least 80% of all sampled students across responding schools.
	<b>Standard 1.13:</b> The final weighted teacher response rate is at least 75% of all sampled teachers across responding schools.
	<b>Standard 1.14:</b> The final weighted sampling unit response rate for any optional cognitive assessment is at least 80% of all sampled students across responding schools.
	<b>Standard 1.16:</b> Unless otherwise agreed upon, the international contractors will draw the school sample for the Main Survey.
	<b>Standard 1.17:</b> Unless otherwise agreed upon, the National Centre will use the sampling contractor's software to draw the student sample, using the list of eligible students provided for each school. Other sampling issues such as undercoverage, poor student listing etc.
	Language of Testing
In all cases the choice of test language(s) in the assessment instruments is made prior to the administration of the test.	
<b>Standard 3.1:</b> PISA participants participating in the PISA2021 Main Survey will have successfully implemented the Field Trial. Unless otherwise agreed upon:	
Field Trial Participation	



Area	Standard
	<p>A Field Trial should occur in an assessment language if that language group represents more than 5% of the target population. For the largest language group among the target population, the Field Trial student sample should be a minimum of 200students per item.</p> <p>For all other assessment languages that apply to at least 5% of the target population, the Field Trial student sample should be a minimum of 100students per item.</p> <p>For additional adjudicated entities, where the assessment language applies to at least 5% of the target population in the entity, the Field Trial student sample should be a minimum of 100students per item.</p>
Adaptation of tests, questionnaires and manuals	<p><b>Standard 4.1:</b> The majority of test items used in previous cycles will be administered unchanged from their previous administration, unless amendments have been made to source versions, or outright errors have been identified in the national versions.</p> <p><b>Standard 4.2:</b> All assessment instruments are equivalent to the source versions. Agreed upon adaptations to the local context are made if needed.</p> <p><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 source versions, outright errors have been identified in the national versions, or a change in the national context calls for an adjustment.</p> <p><b>Standard 4.4:</b> The questionnaire instruments are equivalent to the source versions. Agreed upon adaptations to the local context are made if as needed.</p> <p><b>Standard 4.5:</b> School-level materials are equivalent to the source versions. Agreed upon adaptations to the local context are made as needed.</p>
Translation of assessment instruments, questionnaires, and manuals	<p><b>Standard 5.1:</b> The following documents are translated into the assessment language in order to be linguistically equivalent to the international source versions.</p> <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> <p><b>Standard 5.2:</b> Unless otherwise agreed upon, school-level materials are translated/adapted into the assessment language to make them functionally equivalent to the international source versions.</p>
Testing of national software versions	<p><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.</p> <p><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. Testing results must be submitted to the international contractors so that any errors can be promptly resolved.</p>
Test administration	<p><b>Standard 8.1:</b> All test sessions follow international procedures as specified in the PISA operations manuals, particularly the procedures that relate to:</p> <ul style="list-style-type: none"> <li>• test session timing,</li> <li>• maintaining test conditions,</li> <li>• responding to students' questions,</li> <li>• student tracking, and</li> <li>• assigning assessment materials.</li> </ul> <p><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.</p> <p><b>Standard 8.3:</b> National Centres must not offer rewards or incentives that are related to student achievement in the PISA test to students, teachers, or schools.</p>
Training Support	<p><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.</p> <p><b>Standard 9.2:</b> NPMs or designees shall use the comprehensive training package developed by the contractors and provided on the PISA Portal to train PISA test administrators.</p> <p><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 make results less comparable.</p> <p><b>Standard 9.4:</b> In-person and/or web based test administrator trainings should be conducted by the NPMs or designees, unless a suitable alternative is agreed upon.</p> <p><b>Standard 9.5:</b> PQMs need to successfully complete self-training materials, attend webinars to review and enhance the self-training, and attend the test administrator training, unless otherwise agreed upon.</p>
National Options	<p><b>Standard 10.1:</b> Only national options that are agreed upon between the National Centre and the international contractors are implemented.</p> <p><b>Standard 10.2:</b> Any national option instruments that are not part of the core components 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.</p>

Area	Standard
Security of the material	<p><b>Standard 11.1:</b> PISA materials designated as secure are kept confidential at all times. Secure materials include all test materials, data, and draft materials. In particular:</p> <ul style="list-style-type: none"> <li>• no-one other than approved project staff and participating students during the test session is able to access and view the test materials,</li> <li>• no-one other than approved project staff will have access to secure PISA data and embargoed material, and</li> <li>• formal confidentiality arrangements will be in place for all approved project staff.</li> </ul>
	<p><b>Standard 11.2:</b> Participating schools, students and/or teachers should only receive general information about the test prior to the test session, rather than formal content-specific training. In particular, it is inappropriate to offer formal training sessions to participating students, in order to cover skills or knowledge from PISA test items, with the intention to raise PISA scores.</p>
Quality Monitoring	<p><b>Standard 12.1:</b> PISA Main Survey test administration is monitored using site visits by trained independent quality monitors.</p>
	<p><b>Standard 12.2:</b> Fifteen site visits to observe test administration sessions are conducted in each PISA participating country/economy, and five site visits in each adjudicated region.</p>
	<p><b>Standard 12.3:</b> Test administration sessions that are the subject of a site visit are selected by the international contractors to be representative of a variety of schools in a country/economy.</p>
Printing of Materials	<p><b>Standard 13.1:</b> All paper-based student assessment material will be centrally assembled by the international contractors and must be printed using the final printready file and agreed upon paper and print quality. New countries/entities must submit a printed copy of all Field Trial instruments (booklets and questionnaires) for approval of the printing quality for the Main Survey. The same printing standard must be used for both the Field Trial and the Main Survey.</p>
	<p><b>Standard 13.2:</b> The cover page of all national PISA test paper-based materials used for students and schools must contain all titles and approved logos in a standard format provided in the international version.</p>
	<p><b>Standard 13.3:</b> The layout and pagination of all test paper-based material is the same as in the source versions, unless otherwise agreed upon.</p>
	<p><b>Standard 13.4:</b> The layout and formatting of the paper-based questionnaire material is equivalent to the source versions, with the exception of changes made necessary by national adaptations.</p>
Response Coding	<p><b>Standard 14.1:</b> The coding scheme described in the coding guides is implemented according to instructions from the international contractors' item developers.</p>
	<p><b>Standard 14.3:</b> Both the single and multiple coding procedures must be implemented as specified in the PISA operations manuals (see Note 14.1). These procedures are implemented in all software that countries will be required to use.</p>
	<p><b>Standard 14.4:</b> Coders are recruited and trained following agreed procedures.</p>
	<p><b>Standard 14.2:</b> Representatives from each National Centre attend the international PISA coder training session for both the Field Trial and the Main Survey.</p>
Data Submission	<p><b>Standard 15.1:</b> Each PISA participant submits its data in a single complete database, unless otherwise agreed upon.</p>
	<p><b>Standard 15.2:</b> 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 contractors following specifications in the corresponding operational manuals and international/national record layouts (codebooks). Data are submitted in the DME format.</p>
	<p><b>Standard 15.3:</b> Data for all instruments are submitted. This includes the assessment data, questionnaires data, and tracking data as described in the PISA operations manuals.</p>
	<p><b>Standard 15.4:</b> Unless agreed upon, all data are submitted without recoding any of the original response variables.</p>
	<p><b>Standard 15.5:</b> Each PISA participating country's database is submitted with full documentation as specified in the PISA operations manuals.</p>
Communication with the International Contractors	<p><b>Standard 16.2:</b> 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:</p> <ul style="list-style-type: none"> <li>• Are authorized to respond to queries,</li> <li>• Are able to communicate in English,</li> <li>• Acknowledge receipt of queries within one working day,</li> <li>• 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.</li> </ul>
Schedule for submission of materials	<p><b>Standard 18.1:</b> An agreed upon Translation Plan will be negotiated between each National Centre and the international contractors.</p>
	<p><b>Standard 18.2:</b> The following items are submitted to the international contractors in accordance with agreed timelines:</p> <ul style="list-style-type: none"> <li>• the Translation Plan</li> <li>• a print sample of booklets prior to final printing, for new countries/entities using the paper-based instruments (where this is required, see Standard 13.1),</li> <li>• results from the national checking of adapted computer-based assessment materials and questionnaires,</li> <li>• adaptations to school-level materials,</li> <li>• sampling forms (see Standard 1),</li> <li>• demographic tables,</li> <li>• completed Field Trial and Main Survey Review Forms, and</li> <li>• documents related to PISA Quality Monitors: nomination information, Test Administrator training schedules, translated school-level materials, school contact information, test dates, and</li> </ul>

Area	Standard
	<ul style="list-style-type: none"> <li>• other documents as specified by the PISA operations manuals</li> </ul>
	<b>Standard 18.3:</b> Questionnaire materials are submitted for linguistic verification only after all adaptations have been agreed upon.
	<b>Standard 18.4:</b> All adaptations to those elements of the school-level materials that are required to be functionally equivalent to the source as specified in Standard 5.2, need to be agreed upon.
Management of data	<b>Standard 19.1:</b> 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 eight weeks of the last day of testing for the Main Survey, unless otherwise agreed upon.
	<b>Standard 19.2:</b> National Centres execute data checking procedures as specified in the PISA operations manuals before submitting the database.
	<b>Standard 19.3:</b> National Centres make a data manager available upon submission of the database. The data manager: <ul style="list-style-type: none"> <li>• is authorized to respond to international contractor data queries,</li> <li>• is available for a three-month period immediately after the database is submitted unless otherwise agreed upon,</li> <li>• is able to communicate in English,</li> <li>• is able to respond to international contractor queries within three working days, and</li> <li>• is able to resolve data discrepancies.</li> </ul>
	<b>Standard 19.5:</b> 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.
Archiving of Materials	<b>Standard 19.4:</b> A complete set of PISA paper-based instruments as administered and including any national options, is forwarded to the international contractors on or before the first day of testing. The submission includes the following: <ul style="list-style-type: none"> <li>• hard copies of instruments,</li> <li>• electronic PDF copies of instruments</li> </ul>
	<b>Standard 20.2:</b> The National Project Manager must submit one copy of each of the following translated and adapted Main Survey materials to the international contractors: <ul style="list-style-type: none"> <li>• electronic versions (Word and/or PDF) of all administered Test Instruments, including international and national options</li> <li>• electronic versions (Word and/or PDF) of all administered Questionnaires, including international and national options (paper-based countries only);</li> <li>• electronic versions of the school-level materials; and</li> <li>• electronic versions of the Coding Guides.</li> </ul>
Data Suppression for Privacy Rights	<b>Standard 21.4:</b> Each National Centre must facilitate requests from participants to exercise their data rights. <ul style="list-style-type: none"> <li>• Data access requests will be possible using the raw data from the assessment. No scaled data will be provided in breach of the PISA data embargo.</li> <li>• Data erasure requests will be possible for a limited period before submission to the Contractors. This is to be decided by each National Centre, with two options, up to the submission of ST12 or to upload of student data files to the OECS.</li> <li>• Each National Centre will retain and update a log of completed data requests for data erasure, to facilitate quality control processes. This information must be submitted to the PISA contractors in a timely manner to comply with the requests and for the purpose of data management and sampling processes.</li> </ul>
Meeting Attendance	<b>Standard 23.1:</b> Representatives from each National Centre are required to attend all PISA international meetings including National Project Manager meetings, coder training, and any separate within-school sampling training, and data management training, as necessary. Up to 6 international meetings are planned per cycle.
	<b>Standard 23.2:</b> Representatives from each National Centre who attend international meetings must be able to work and communicate in English.
Invariant Item Parameters	<i>TAG Memo Dec 2021:</i> For each major language of assessment within a participating country/economy, over two-thirds of items per domain are expected to be invariant from the international item parameters for the Field Test and the Main Survey.

Note: Albeit not a Technical Standard *per se*, a significant proportion of common items are essential for the linking of PISA national versions among jurisdictions and thus central for the comparability of assessment results. The PISA Technical Advisory Group (TAG) has fixed a threshold for invariant item parameters that will be incorporated into PISA technical documents in future cycles.

Source: PISA 2022 Technical Standards (Annex I)

# 17 Proficiency Scale Construction for the Core Domains

## Introduction

This chapter discusses the methodology used to develop the PISA Mathematics reporting scales. These describe levels of proficiency in the domain and presents the outcomes of the development process for mathematics literacy, the major domain in the PISA 2022 assessment.

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/economy 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.

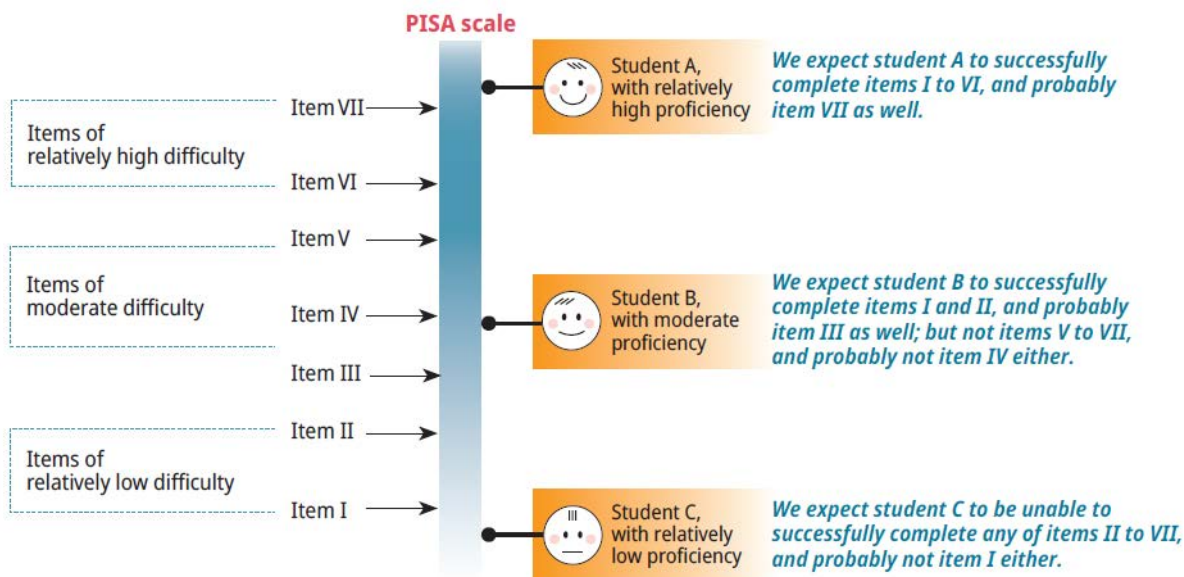
The PISA test design makes it necessary to use techniques of modern item response modelling to both, estimate the ability of all students taking the PISA assessment and the statistical characteristics of all PISA items. These techniques are described in Chapter 11 [*Scaling PISA Data*].

The PISA data are collected using a rotated matrix test design in which students take different but overlapping sets of items. 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 of the items 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, 2023<sup>[1]</sup>; 2023<sup>[2]</sup>; 2023<sup>[3]</sup>). 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 17.1 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. Seven 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.

**Figure 17.1. Simplified relationship between items and students on a proficiency scale**



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/economies.

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 PISA 2000 and was revised in the lead-up to each of the subsequent surveys. The same basic methodology has again been used to develop proficiency descriptions for the core domains of PISA 2022, even though, like in the PISA 2015 and PISA 2018 cycles, a more general statistical model to describe the items was used in the scaling procedure compared to PISA cycles before 2015.

The proficiency descriptions that had been developed for the science domain in PISA 2015, for the reading domain in PISA 2018, and for financial literacy in 2012 were used again to report the results of PISA 2022. The proficiency descriptors for creative thinking, the innovative domain for PISA 2022, are entirely new and these are described in the next chapter of this Technical Report.

Reporting for mathematics, the major domain in PISA 2022, was linked back to the 2012 proficiency scale and was based on the detailed proficiency level descriptions developed in 2012, the last PISA cycle in which mathematics was the major domain. These proficiency level descriptors were revised based on PISA 2022 data in order to incorporate the new aspects of the mathematics framework and the performance of the new items, including the reasoning and interactive items.

The mathematics expert group (MEG) worked with the Core A contractor (ETS) to revise the sets of described proficiency scales and subscales for PISA mathematics. Similarly, the Creative Thinking Expert

Group (CTEG) worked with Core B3 contractor (ACT) to develop the described proficiency scale for that domain. More detail on the development of the Creative Thinking scale is given in Chapter 18.

## Development of the PISA scales

The development of described proficiency scales for PISA has been carried out through a process that typically involves several 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 PISA 2022. 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 administration. The following description of the development process focuses on the development of described proficiency scales for mathematics.

### ***Classification of items***

As part of new item development for mathematics, test developers classified all items based on the specifications provided in the framework. Item classifications for the trend mathematics items were also revised to reflect the PISA 2022 assessment framework. All trend classifications were reviewed by the MEG and revised as needed.

### ***Defining the overall proficiency scale***

Using Main Survey data with preliminary student weights, the mathematics expert group met over several days and reviewed representative items, particularly those that were classified as representing the new reasoning process scale or having an interactive component (e.g. a spreadsheet or data simulator) and discussing key characteristics that differentiated performance along the proficiency scale. Following this meeting, the descriptors for each level in the overall proficiency scale were refined and finalised.

### ***Identifying possible subscales***

For each major domain assessed in PISA, reporting includes an overall proficiency scale based on the combined results for all items within that domain. In addition, the assessment 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.

In the case of mathematics, a single mathematical scale was developed for PISA 2000. With the additional data available in PISA 2003, when mathematics was the major test domain, subscales based on the four overarching subdomains – *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 PISA 2012 could be incorporated into the reporting for the cycle. The outcome of these considerations was, first, 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. Second, the expert group recommended that three process-based subscales should be

reported. These included: *formulating situations mathematically* (or “formulate”), *employing mathematical concepts, facts, and procedures* (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 just “uncertainty”), were also reported. For PISA 2015 and 2018, where mathematics was once more the minor domain, only a single scale representing overall proficiency in the mathematics domain was reported.

For the PISA 2022 cycle, the MEG decided that additional proficiency scales should be reported for the four mathematical processes (i.e. *mathematical reasoning; formulate; employ; and interpret*). Since the last three processes were part of the domain in previous cycles, proficiency scales already existed and just needed to be updated based on the new items classified to each process. However, mathematical reasoning was “new” this cycle as a separate process scale, so that a proficiency scale needed to be fully developed. As part of their work updating the mathematics framework, the MEG developed a range of actions that students would be expected to perform for each of the mathematical processes. These actions represented a hierarchy of “demands” that the items make of the students in order to solve a problem and were designed to span the proficiency scale. These lists of actions proved useful during the item development phase and when updating/writing the proficiency scales.

### **Scales in the minor domains**

For science, 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 PISA 2018 and PISA 2022, only a single scale representing overall proficiency in the science domain is reported.

For reading, which was the major domain in PISA 2018, work on identifying possible subscales began with a review of the subscales used in PISA 2009, when reading was also a major domain. In PISA 2009, volume I of the *PISA 2009 Results* included an overall reading scale and descriptions of subscales that described the types of reading tasks or “cognitive aspects”: access and retrieve, integrate and interpret and reflect and evaluate and subscales based on the form of reading material: continuous texts and non-continuous texts (OECD, 2010<sup>[4]</sup>). For digital reading, a separate, single scale was developed based on the digital reading assessment items administered in 19 countries/economies in PISA 2009, as an international option (OECD, 2011<sup>[5]</sup>). In PISA 2012, when reading was a minor domain, a single print reading scale was reported, along with a single digital reading scale. For PISA 2018, the reading expert group decided the former distinction of “cognitive aspects” should be updated to “cognitive processes”. This terminology better connects the PISA 2018 assessment framework with the literature on reading psychology and better reflects the actual skills and proficiencies assessed. The subscales that correspond to the ways students interact and process text were updated to the following: locate information, understand, and evaluate and reflect. The former subscales that were based on the form of reading material are not included in PISA 2018. Instead, scales are included corresponding to using a single unit of text or multiple units of texts for answering the questions. For PISA 2022, only a single scale representing overall proficiency in reading was reported.

For creative thinking, the innovative domain in PISA 2022, a proficiency description on a single overall reporting scale was developed and is described in the next Chapter. The optional assessment of financial literacy used the same proficiency description from PISA 2015 and PISA 2018.

## Defining the proficiency levels

The proficiency levels for each of the PISA domains were defined when each was first introduced as a major domain. The goal of that process was to decide how to divide up the proficiency continuum into levels that might be more interpretable. 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 and student proficiency and item characteristics is probabilistic. That is, there is some probability that a particular student can correctly solve a 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 student proficiency.

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 same basic methodology has again been used to develop proficiency descriptions for PISA 2022.

Defining proficiency levels for PISA 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 that underpin each assessment domain. This produces descriptions of increasing proficiency that reflect observations of student performance and a detailed analysis of the cognitive demands of PISA assessment items. The second phase involves 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 the PISA assessment 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 said 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 clothing and portions by size (i.e. small, medium, large, extra-large, etc.).

The approach adopted since 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 border of a level would complete at least 50% of items 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 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 range. 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.

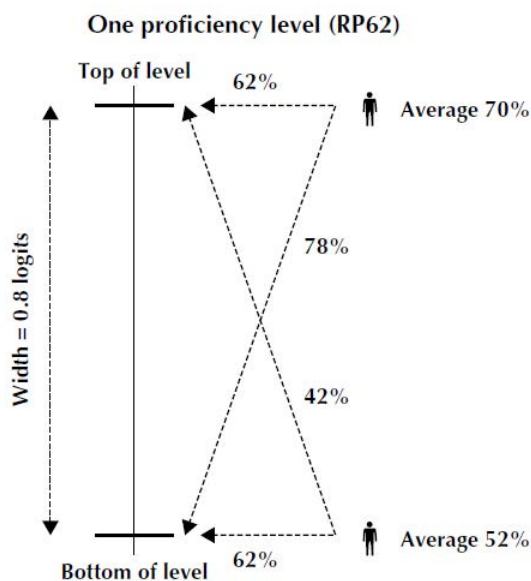
Thus, a consistent approach should be taken to defining levels for the different scales. Their range may not be exactly the same for the proficiency scales in different assessment domains, but the same kind of interpretation should be possible for each scale that is developed. This approach links the following three variables:

- 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 17.2 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 assessment 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. 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 17.2. Calculating the RP values used to define PISA proficiency levels**



In PISA we have 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 range as the other described levels. Student performance below this level is lower than that which PISA can reliably assess and, more importantly, describe.

## Reporting PISA results for Mathematics

In this section, the ways in which levels of mathematics are defined, described and reported will be discussed. This will be illustrated using a subset of released new mathematics items from PISA 2022.

### *Building an item map for mathematics*

The data from the PISA mathematics assessment were analysed to estimate a set of item characteristics for the 234 items included in the Main Survey. During the process of item development, each item was classified to reflect the content area and mathematical process it required. Following data analysis, the items were associated with their difficulty. Annex Table 17.A.1 shows an item map, which includes information for a set of new mathematics units released after the PISA 2022 Main Survey. Each row in Annex Table 17.A.1 represents a level on the mathematics proficiency scale. The selected items have been ordered according to their difficulty, with the most difficult at the top, and the least difficult at the bottom of the table. The difficulty estimate for each item expressed in the reporting scale is given in the rightmost column. For items with a partial-credit response category, the item is listed twice to show the level for a full-credit response and the level for a partial-credit response. Partial-credit responses are listed in italicised font. Note that four new mathematics units were also released after the Field Trial, but those units are not included here because there were no estimates of item difficulty.

## Defining levels of mathematical literacy

The reporting approach used by the OECD has been defined in previous PISA cycles 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/economies. A similar approach has been used here to analyse and report PISA 2022 assessment outcomes for mathematics.

Since the PISA 2000 assessment, results have been reported on a scale with a mean of 500 and a standard deviation of 100. The metric has been set using the participating OECD countries at the time when the subject was the major domain for the first time. In PISA 2012, the last time mathematics was the major domain, the scale consisted of Levels 1 through 6. Starting with the PISA for Development (PISA-D) assessment, Level 1 on the mathematics scale was split into Levels 1c, 1b, and 1a, with Level 1a corresponding to what had previously just been Level 1. This was done to further describe what students at the lower levels of proficiency can do. The level definitions on the PISA mathematical literacy scale are given in Annex Table 17.A.2.

Information about the items in each level is used to develop summary descriptions of the kinds of mathematical skills and abilities associated with different levels of proficiency. These summary descriptions can then be used to encapsulate typical mathematical proficiency of students associated with each level. As a set, they describe a progression in mathematical ability.

For PISA 2022, there was already a set of proficiency level descriptors upon which to build. The new items that were developed for PISA 2022 were considered in relation to the existing level descriptions. Annex Table 17.A.3 presents the updated description for the overall mathematical literacy scale Annex Table 17.A.4 Annex Table 17.A.5 and Annex Table 17.A.6 present updated descriptions for each process (i.e. Formulate, Employ, and Interpret, respectively) that was part of the mathematical problem-solving model also used in PISA 2012.

Annex Table 17.A.7 presents a description of the mathematical reasoning process, which for PISA 2022 was treated as separate process.

Annex Table 17.A.8, Annex Table 17.A.9, Annex Table 17.A.9. and Annex Table 17.A.11 present updated descriptions for each content that was part of the mathematics assessment in PISA 2022.

## Cutpoints defining proficiency levels for Reading, Science and Financial Literacy in PISA 2022

Annex Table 17.A.12, Annex Table 17.A.13 and Annex Table 17.A.14 present the cut points used to assign items and students to a proficiency level for the minor domains of reading and science, as well as for the financial literacy domain. As with the mathematics cut points, values in the table are the lower bound for the corresponding level. For example, in the reading scale, Level 6 begins with 698.32. Level 5 begins with 625.61 and ends just below 698.32, where Level 6 begins. Below Level 1c are those with values lower than 189.33. In other words, those reaching a level are those with a score or difficulty at or above the given cut point. This same interpretation applies to all proficiency scales used in PISA.

## References

- OECD (2023), *PISA 2021 Creative Thinking Framework*, PISA, OECD Publishing, Paris, [2]  
<https://www.oecd.org/pisa/publications/PISA-2021-Creative-Thinking-Framework.pdf>.
- OECD (2023), *PISA 2021 Financial Literacy Framework*, PISA, OECD Publishing, Paris, [3]  
<https://www.oecd.org/pisa/sitedocument/PISA-2021-Financial-Literacy-Framework.pdf>.
- OECD (2023), *PISA 2021 Mathematics Framework*, PISA, OECD Publishing, Paris, [1]  
<https://www.oecd.org/pisa/sitedocument/PISA-2021-mathematics-framework.pdf>.
- OECD (2011), *PISA 2009 Results: Students On Line: Digital Technologies and Performance (Volume VI)*, PISA, OECD Publishing, Paris, [5]  
<https://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)*, PISA, OECD Publishing, Paris, [4]  
<https://doi.org/10.1787/9789264091450-en>.

## Annex 17.A. PISA Mathematics reporting scales

Annex Table 17.A.1. A map for released mathematics items

Level	Cut point	Item	Item Difficulty
6	669.30	Forested Area (CMA161Q03) – Full Credit	840
		Forested Area (CMA161Q04)	739
		Points (CMA156Q01) – Full Credit	672
5	606.99	Forested Area (CMA161Q02)	647
		Points (CMA156Q01) – Partial Credit	642
		Forested Area (CMA161Q01) – Full Credit	636
		Triangular Pattern (CMA150Q03) – Full Credit	620
		Forested Area (CMA161Q03) – Partial Credit	617
4	544.68	Forested Area (CMA161Q01) – Partial Credit	575
		Triangular Pattern (CMA150Q03) – Partial Credit	545
3	482.38	Solar System (CMA123Q01) – Full Credit ( <i>Partial Credit</i> )	514 (503)
2	420.07	Triangular Pattern (CMA150Q02)	448
		Solar System (CMA123Q02)	430
1a	357.77	Triangular Pattern (CMA150Q01)	411
1b	295.47	There were no released items at this level	
1c	233.17	There were no released items at this level	

Annex Table 17.A.2. Mathematical literacy performance band definitions on the PISA scale

Level	Score points on the PISA Scale
6	At or above 669.30
5	At or above 606.99 but less than 669.30
4	At or above 544.68 but less than 606.99
3	At or above 482.38 but less than 544.68
2	At or above 420.07 but less than 482.38
1a	At or above 357.77 but less than 420.07
1b	At or above 295.47 but less than 357.77
1c	At or above 233.17 but less than 295.47

Annex Table 17.A.3. Summary descriptions of the proficiency levels on the Mathematical Literacy scale

Level	What students can typically do
6	At Level 6, students can work through abstract problems and demonstrate creativity and flexible thinking to develop solutions. For example, they can recognise when a procedure that is not specified in a task can be applied in a non-standard context or when demonstrating a deeper understanding of a mathematical concept is necessary as part of a justification. They can link different information sources and representations, including effectively using simulations or spreadsheets as part of their solution. Students at this level are capable of critical thinking and have a mastery of symbolic and formal mathematical operations and relationships that they use to clearly

Level	What students can typically do
	communicate their reasoning. They can reflect on the appropriateness of their actions with respect to their solution and the original situation.
5	At Level 5, students can develop and work with models for complex situations, identifying or imposing constraints, and specifying assumptions. They can apply systematic, well-planned problem-solving strategies for dealing with more challenging tasks, such as deciding how to develop an experiment, designing an optimal procedure, or working with more complex visualisations that are not given in the task. Students demonstrate an increased ability to solve problems whose solutions often require incorporating mathematical knowledge that is not explicitly stated in the task. Students at this level reflect on their work and consider mathematical results with respect to the real-world context.
4	At Level 4, students can work effectively with explicit models for complex concrete situations, sometimes involving two variables, as well as demonstrate an ability to work with undefined models that they derive using a more sophisticated computational-thinking approach. Students at this level begin to engage with aspects of critical thinking, such as evaluating the reasonableness of a result by making qualitative judgements when computations are not possible from the given information. They can select and integrate different representations of information, including symbolic or graphical, linking them directly to aspects of real-world situations. At this level, students can also construct and communicate explanations and arguments based on their interpretations, reasoning, and methodology.
3	At Level 3, students can devise solution strategies, including strategies that require sequential decision-making or flexibility in understanding of familiar concepts. At this level, students begin using computational-thinking skills to develop their solution strategy. They are able to solve tasks that require performing several different but routine calculations that are not all clearly defined in the problem statement. They can use spatial visualisation as part of a solution strategy or determine how to use a simulation to gather data appropriate for the task. Students at this level can interpret and use representations based on different information sources and reason directly from them, including conditional decision-making using a two-way table. They typically show some ability to handle percentages, fractions and decimal numbers, and to work with proportional relationships.
2	At Level 2, students can recognise situations where they need to design simple strategies to solve problems, including running straightforward simulations involving one variable as part of their solution strategy. They can extract relevant information from one or more sources that use slightly more complex modes of representation, such as two-way tables, charts, or two-dimensional representations of three-dimensional objects. Students at this level demonstrate a basic understanding of functional relationships and can solve problems involving simple ratios. They are capable of making literal interpretations of results.
1a	At Level 1a, students can answer questions involving simple contexts where all information needed is present, and the questions are clearly defined. Information may be presented in a variety of simple formats and students may need to work with two sources simultaneously to extract relevant information. They are able to carry out simple, routine procedures according to direct instructions in explicit situations, which may sometimes require multiple iterations of a routine procedure to solve a problem. They can perform actions that are obvious or that require very minimal synthesis of information, but in all instances the actions follow clearly from the given stimuli. Students at this level can employ basic algorithms, formulae, procedures, or conventions to solve problems that most often involve whole numbers.
1b	At Level 1b, students can respond to questions involving easy to understand contexts where all information needed is clearly given in a simple representation (i.e. tabular or graphic) and, as necessary, recognise when some information is extraneous and can be ignored with respect to the specific question being asked. They are able to perform simple calculations with whole numbers, which follow from clearly prescribed instructions, defined in short, syntactically simple text.
1c	At Level 1c, students can respond to questions involving easy to understand contexts where all relevant information is clearly given in a simple, familiar format (for example, a small table or picture) and defined in a very short, syntactically simple text. They are able to follow a clear instruction describing a single step or operation.

#### Annex Table 17.A.4. Summary descriptions of the proficiency levels on the Formulating situations mathematically scale

Level	What students can typically do
6	Students at Level 6 can typically apply a wide variety of mathematical content knowledge to transform and represent information from a broad variety of contexts into a mathematical form amenable to analysis. At this level, students can formulate and solve complex real-world problems involving significant modelling steps and extended calculations, such as applying their geometric knowledge to irregular shapes, inferring relevant parameters of a large data set, or analysing an experiment to recognise the mathematical relationship between objects. Students at level 6 are able to identify the relationship between the key components of a problem and to develop algebraic formulations that accurately represent them.
5	At level 5, students show an ability to use their understanding across a range of mathematical areas to transform information or data from a problem context into mathematical form, sometimes involving two or more variables. They are able to recognise a situation where statistical counting techniques can be applied or formulate inequalities based on given conditions. Students are able to manipulate relatively large data sets by determining appropriate mathematical operations to perform using a spreadsheet tool. They are able to analyse more complex geometric figures, for example, by recognising the relationship between the properties of a compound figure and the properties of individual shapes that comprise the compound figure. Students at this level can formulate a process to solve a problem where some of the information used is given as a range instead of a single value or when information is not given explicitly in the task.
4	At Level 4, students are able to solve complex problems in a variety of contexts that may require designing a sequence of steps to reach the solution. They also recognise when a single process, repeated iteratively, can lead to the solution. Students are able to run simulations

Level	What students can typically do
	to identify the underlying relationship between two or more variables. They can determine probabilities from data presented in two-way tables. Students at this level can also formulate linear algebraic expressions of relatively simple contexts involving one constraint, recognise an application of a known procedure from a data table and use that procedure to determine missing values, or formulate a method to compare information, such as the prices of several sale items. They can work with more complex geometric models of practical situations which contain all the relevant information needed for formulating the solution.
3	At Level 3, students can identify and extract information from a variety of sources, including text, geometric models, tables, and diagrams, where all necessary information is provided. They can identify basic mathematical concepts relevant for the model or identify how to transform information given in a diagram to data that can be input into a simulation. Students at this level are able to solve problems by recognising situations in which quantities are related proportionally or by performing a computation using a percentage in real-life contexts such as medical testing or ticket sales. They are able to solve simple multi-step problems where the sequence of steps needs to be determined, and each step requires translating some of the given information into a form that can be operated on mathematically.
2	At this level, students can understand clearly formulated instructions and information about simple processes and tasks in order to express them in a mathematical form. They can determine a rule used in a simple pattern, and then use that rule to extend the pattern to the next term. They are able to use information presented in tables or diagrams to identify or build a simple model of a practical situation. For example, they can revise a given formula to determine the number of seats in any row of a theatre. Students at this level are able to translate descriptions of situations to be operated on mathematically that first require identifying information relevant to the particular task. At this level, students begin to formulate situations involving non-integer quantities, provided all necessary information is given in the task.
1a	At this level, students can recognise an explicit model of a contextual situation from a list or translate a short verbal description so that it can be operated on using basic mathematical tools. Students at this level are able to work with simple models involving one operation and at most two variables. For example, they can select the appropriate model that represents the total number of items that can be produced based on a production rate. Students at this level are capable of formulating situations that involve whole numbers and where all relevant information is given.
1b	<i>There were no items to describe this level on the scale.</i>
1c	<i>There were no items to describe this level on the scale.</i>

**Annex Table 17.A.5. Summary descriptions of the proficiency levels on the Employing mathematical concepts, facts, and procedures scale**

Level	What students can typically do
6	Students at Level 6 are typically able to employ a strong repertoire of knowledge and procedural skills in a wide range of mathematical areas. They can solve problems involving several stages or a problem that does not have a well-defined solution method, such as computing the area of an irregularly shaped figure. They demonstrate an understanding of statistical data, and can apply that understanding, for example, to determine the probability of different events. Students at this level can observe regularities in information and use that to determine algorithms to apply to a situation. At Level 6, students' work is consistently precise and reflects a strong ability to work with different data formats and representations.
5	Students at Level 5 can employ a broader range of knowledge and skills to solve problems. They can sensibly link information in graphical and diagrammatic formats to textual information. Students can reason proportionally to find a unit rate or understand and apply the meaning of a concept to extract relevant information from a table to solve a problem. At this level, they can devise a strategy to extrapolate from a sample or to determine which of two savings options would be better in a situation involving variously priced items. Students demonstrate the ability to solve problems that require converting between units or working with constraints and can provide mathematical or conceptual arguments to support their results. They also demonstrate proficiency working with percentages and ratios.
4	At Level 4, students show an understanding of the context and can recognise efficient strategies for solving problems. For example, they can typically identify relevant data and information from contextual material and use it to perform such tasks as, calculating distances from a map, analysing a model based on percentages, or comparing the results from two different formulae to compute the same measure. They are able to determine how a rating system was used to support a claim or evaluate several construction designs to rank order them based on a given criterion. At this level, students can estimate values from a graph and use them to solve a problem or analyse statements relating quantities expressed in different numerical formats. They demonstrate an ability to work with ratios or problems that require a series of steps be performed in a specific order.
3	Students at Level 3 demonstrate more flexibility in devising and implementing solution strategies for problems that can be solved in a variety of ways. They are able to solve problems where the information given in the task must first be analysed to determine which of a given set of processes should be implemented, such as determining a fine for exceeding a speed limit based on different driving speeds or a model for computing charges for water-usage. At this level, students are able to use the basic properties of angles to solve a geometric problem or are able to translate between graphical and tabular representations of the same data. Students show an ability to approximate a final solution from interim results or to recognise how a given constraint affects the conclusion. They can work with percentages, fractions, decimal numbers, proportional relationships, and simple non-linear contexts.
2	Students at Level 2 show an ability to work with given models in flexible ways, such as identifying the relevant information to input or manipulating information to make it amenable to use in the model (including models with multiple inputs or tasks that require using a calculator tool specific to the context). They are also able to determine the input when given the output. Students can apply familiar geometric concepts to analyse a spatial pattern. At this level, students show an understanding of place value in decimal numbers and can use that understanding to compare numbers presented in a familiar context. They can apply a known procedure that first requires

Level	What students can typically do
	understanding a data table to extract the necessary information. Students are able to solve simple problems using proportional reasoning and work with ratios.
1a	Students at Level 1a can solve well-defined problems that require minimal decisions. For example, they can make direct inferences from textual information that points to an obvious strategy to solve a given problem, particularly where the mathematical procedures are one- or two-step arithmetic operations with whole numbers or require application of a familiar procedure. Students are able to extract information presented in a variety of formats, such as advertisements, simple pie charts, diagrams, or tables, which contain all the needed information to solve a problem. At this level, students can compute simple percentages, recognise when quantities are related proportionally, find the total area of a standard region, or determine a cost saving.
1b	At Level 1b, students can employ straightforward, one-step procedures that are clearly defined in the task, and where all information is presented in simple tabular format. For example, they are able to determine the winner of a tournament given the criterion for winning or locate information in a table based on a set of conditions.
1c	<i>There were no items to describe this level on the scale.</i>

**Annex Table 17.A.6. Summary descriptions of the proficiency levels on the Interpreting, applying and evaluating mathematical outcomes scale**

Level	What students can typically do
6	At Level 6, students are able to link multiple complex mathematical representations in an analytical way to identify and extract data and information that enables conceptual and contextual questions to be answered. Students at this level demonstrate creativity in order to evaluate claims or interpret solutions to problems that require greater insight to solve, such as using a simulation to determine a design that satisfies several conditions. They are able to interpret data sets with multiple variables that typically require having to perform two or more operations before being able to evaluate a set of given claims related to the data set. Students can recognise different possible subdivisions of an irregular shape based on interpreting a list of geometric properties of the irregular shape. At this level, students can readily interpret or evaluate percentages, frequency distributions, and statistical measures, such as means and medians, in a variety of contexts.
5	At Level 5, students demonstrate the ability to interpret complex situations that require analyses of the underlying mathematics and can apply their understanding of mathematical concepts to real-world situations to make judgements on the reasonableness of claims or results. For example, students can explain why a possible mathematical model does not fit the real-world context. They can interpret experimental results and devise a method for comparing and ranking the results based on a given criterion. At this level, students can evaluate statistical statements based on means or product ratings presented in multiple formats, or they can manipulate a data set so that the presentation facilitates interpretation of the provided information.
4	At Level 4, students are able to interpret and evaluate situations or outcomes that typically involve satisfying multiple conditions, in a range of real-world contexts. They are able to interpret simple statistical or probabilistic statements from data presented in tables or charts in such contexts as fitness levels or genetics. Students at this level are able to interpret experimental results to infer a relationship between two variables in order to evaluate a claim or explain how the computational result of an experiment relates to a given set of specifications. They can determine if a solution is compatible with a particular context or recognise how different adjustments to an algorithm affect the results. At this level, students also are able to approach problems where their interpretation of the given information or model can influence the solution strategy they choose for the task.
3	Students at Level 3 show an ability to reflect on an outcome, or the process used to reach an outcome, in more complex contexts. For example, they can interpret an algebraic model of a design plan to determine what quantity a variable in the model represents or manipulate a set of data using a spreadsheet tool to analyse claims related to energy usage or changes in population data. Students are able to use simulation results to determine a relationship between two contextual variables or explain if a conjecture about a simple algorithm is true. Students demonstrate spatial reasoning by translating between two- and three-dimensional representations of solids or by understanding how properties of geometric figures are related. At this level, students can analyse relatively unfamiliar data presentations to support their conclusions or interpret solutions of non-integer values or ratios with respect to real-world contexts.
2	At Level 2, students can link conceptual and contextual elements of the problem to the mathematics in order to solve problems in a variety of real-world contexts where the information is presented clearly. Students are able to evaluate outcomes, often without having to perform calculations, such as determining the angle measures of an object based on interpreting a description of its properties. They can interpret context-specific language into simple mathematical relationships, sometimes involving one or two constraints, or understand how relationships presented in graphical formats relate to the context, such as a graph of distance versus time. At this level, students can run simulations and interpret the results with respect to the conditions of the task involving one variable.
1a	At Level 1a, students are able to locate and utilise information in order to make sense of the context. They can interpret information that requires relating two simple data sources, such as tables. For example, they can relate information in one table showing how points are awarded to another table of match outcomes to solve a problem in a familiar context or to understand how data from one source is represented in another source. Students at this level can also recognise when some of the given information can be ignored with respect to the specific task.
1b	At Level 1b, students are able to interpret contextual information presented in one of a variety of formats, such as two-way tables or work schedules. They demonstrate an ability to process the information given basic constraints imposed by the task, such as determining which rule from a table to apply or when to plan an event.



Level	What students can typically do
1c	Students at Level 1c can interpret information from real-world contexts presented in simple diagrams or tables and then use that information to solve well-defined problems involving a single operation with whole numbers or straightforward comparisons.

### Annex Table 17.A.7. Summary descriptions of the proficiency levels on the Mathematical reasoning scale

Level	What students can typically do
6	At Level 6, students use deductive and inductive reasoning to devise strategies to solve real-world problems that require inference and creativity to recognise the mathematical nature of the task. Tasks at this level are often presented abstractly and require reasoning to recognise how the context-specific language can be transformed into known mathematical concepts or procedures, which underlies making the mathematical context suitable for analysis. Students can solve problems that require visualising a nonstandard geometric model not explicitly shown or described in the task or that require a solid understanding of known algorithms. For example, they can transform given information to construct a visual model to represent a situation or they can use the definition of a procedure for computing a statistical measure to justify if a mathematical result is possible without having numerical values to manipulate. At this level, they use reasoning to critique the limits of a model, such as identifying if a model can or cannot be used in a particular situation, which is necessary for being able to interpret/evaluate the mathematical outcome in context. Students also use reasoning to construct mathematical arguments based on logic and contradictions, such as justifying if a conclusion can be made from a given data set or developing a counterexample in response to a claim.
5	At Level 5, students can recognise structure in problem situations that can be solved using an algorithmic approach. Students use computational thinking to design an optimal procedure, such as programming a sequence of commands, and then reflect on the solution to determine if it meets the given constraints. They can analyse situations and recognise how a known procedure or set of procedures can be applied as a way to justify, for example, if an object can fit into a particular space or if a plan for a geometric design is possible. At this level, they can determine how to develop an experiment and run simulations to collect data necessary for evaluating a context. Students can identify a counterexample or analyse a rule used in a pattern as a way to support a mathematical argument. Students also use reasoning to develop solution strategies by identifying which elements of a model vary and which are invariant.
4	At Level 4, students demonstrate reasoning ability by reflecting on solutions to explain mathematical concepts in real-world contexts. They can evaluate the reasonableness of a claim and provide mathematical justifications to either support or refute the claim, such as recognising how to apply a common procedure in a novel context or determining how to interpret data or information presented in articles, tables, or phone apps. At this level, students can use their understanding of arithmetic and algebraic properties to analyse how manipulating the variables in a model or the steps in a procedure will help explain the real-world results, or they can develop a model to derive a relationship between the variables used in an equation. Students can identify more complex geometric relationships from images of shapes or descriptions of their properties. They are able to reason inductively from sample results to inform decision making or reason about the likelihood of various outcomes related to a probability context.
3	At Level 3, students can apply reasoning by utilising definitions and making judgements necessary for transforming conceptual and contextual situations into mathematical problems. Students at this level can evaluate a claim based on devising simple strategies to connect the underlying mathematics with the context. They are able to solve problems that require making minimal assumptions, such as recognising the relative size of a region from a diagram or comparing graphs of population data. Students can reason about properties in a description of a geometric model to determine a simple algebraic relationship. At this level, they can also apply reasoning to solve problems involving familiar concepts presented in nonstandard ways, such as race results or statistical measures represented graphically on a coordinate plane.
2	At Level 2, students are able to use reasoning to infer relationships between conceptual and contextual elements in a problem or to devise a straight-forward strategy for evaluating a claim. For example, they can order objects by recognising how the size of various objects relates to distance traveled or how to use given assumptions to compare two rate plans with varying prices. Students at this level can also use spatial reasoning, when provided with a model or diagram, to recognise an alternate representation of an image or to analyse simple geometric properties of the model.
1a	At Level 1a, students use reasoning to draw conclusions based on their understanding of simple mathematical concepts, such as evaluating the likelihood of an outcome in a familiar probability context.
1b	<i>There were no items to describe this level on the scale.</i>
1c	<i>There were no items to describe this level on the scale.</i>

### Annex Table 17.A.8. Summary descriptions of the proficiency levels on the mathematical content subscale: Change and relationships

Level	What students can typically do
6	At Level 6, students use significant insight, abstract reasoning and argumentation skills and technical knowledge and conventions to solve problems involving relationships among variables and to generalise mathematical solutions to complex real-world problems. They are able to create and use an algebraic model of a functional relationship incorporating multiple quantities. They apply deep geometrical

	insight to work with complex patterns. And they are typically able to use complex proportional reasoning, and complex calculations with percentage to explore quantitative relationships and change.
5	At Level 5, students solve problems by using algebraic and other formal mathematical models, including in scientific contexts. They are typically able to use complex and multi-step problem-solving skills, and to reflect on and communicate reasoning and arguments, for example in evaluating and using a formula to predict the quantitative effect of change in one variable on another. They are able to use complex proportional reasoning, for example to work with rates, and they are generally able to work competently with formulae and with expressions including inequalities.
4	Students at Level 4 are typically able to understand and work with multiple representations, including algebraic models of real-world situations. They can reason about simple functional relationships between variables, going beyond individual data points to identifying simple underlying patterns. They typically employ some flexibility in interpretation and reasoning about functional relationships (for example in exploring distance-time-speed relationships) and are able to modify a functional model or graph to fit a specified change to the situation; and they are able to communicate the resulting explanations and arguments.
3	At Level 3, students can typically solve problems that involve working with information from two related representations (text, graph, table, formulae), requiring some interpretation, and using reasoning in familiar contexts. They show some ability to communicate their arguments. Students at this level can typically make a straight-forward modification to a given functional model to fit a new situation; and they use a range of calculation procedures to solve problems, including ordering data, time difference calculations, substitution of values into a formula, or linear interpolation.
2	Students at Level 2 are typically able to locate relevant information on a relationship from data provided in a table or graph and make direct comparisons, for example to match given graphs to a specified change process. They can reason about the basic meaning of simple relationships expressed in text or numeric form by linking text with a single representation of a relationship (graph, table, simple formula), and can correctly substitute numbers into simple formulae, sometimes expressed in words. At this level, student can use interpretation and reasoning skills in a straight-forward context involving linked quantities.
1a	Students at Level 1a are typically able to evaluate single given statements about a relationship expressed clearly and directly in a formula, table, or graph. Their ability to reason about relationships, and change in those relationships, is limited to simple expressions and to those located in familiar situations, such as contexts involving unit rates. They may apply simple calculations needed to solve problems related to clearly expressed relationships.
1b	<i>There were no items in the PISA 2022 Mathematics assessment to describe this level on the scale.</i>
1c	<i>There were no items in the PISA 2022 Mathematics assessment to describe this level on the scale.</i>

### Annex Table 17.A.9. Summary descriptions of the proficiency levels on the mathematical content subscale: Quantity

Level	What students can typically do
6	At Level 6 and above, students conceptualise and work with models of complex quantitative processes and relationships; devise strategies for solving problems; formulate conclusions, arguments and precise explanations; interpret and understand complex information, and link multiple complex information sources; interpret graphical information and apply reasoning to identify, model and apply a numeric pattern. They are able to analyse and evaluate interpretive statements based on data provided; work with formal and symbolic expressions; plan and implement sequential calculations in complex and unfamiliar contexts, including working with large numbers, for example to perform a sequence of currency conversions, entering values correctly and rounding results. Students at this level work accurately with decimal fractions; they use advanced reasoning concerning proportions, geometric representations of quantities, combinatorics and integer number relationships; and they interpret and understand formal expressions of relationships among numbers, including in a scientific context.
5	At Level 5, students are able to formulate comparison models and compare outcomes to determine best price; interpret complex information about real-world situations (including graphs, drawings and complex tables, for example two graphs using different scales); they are able to generate data for two variables and evaluate propositions about the relationship between them. Students are able to communicate reasoning and argument; recognise the significance of numbers to draw inferences; provide a written argument evaluating a proposition based on data provided. They can make an estimation using daily life knowledge; calculate relative and/or absolute change; calculate an average; calculate relative and/or absolute difference, including percentage difference, given raw difference data; and they can convert units (for example calculations involving areas in different units).
4	At Level 4, students are typically able to interpret complex instructions and situations; relate text-based numerical information to a graphic representation; identify and use quantitative information from multiple sources; deduce system rules from unfamiliar representations; formulate a simple numeric model; set up comparison models; and explain their results. They are typically able to carry out accurate and more complex or repeated calculations, such as adding 13 given times in hour/minute format; carry out time calculations using given data on distance and speed of a journey; perform simple division of large multiples in context; carry out calculations involving a sequence of steps and accurately apply a given numeric algorithm involving a number of steps. Students at this level can perform calculations involving proportional reasoning, divisibility or percentages in simple models of complex situations.
3	At Level 3, students typically use basic problem-solving processes, including devising a simple strategy to test scenarios, understand and work with given constraints, use trial and error, and use simple reasoning in familiar contexts. At this level students typically can interpret a text description of a sequential calculation process, and correctly implement the process; identify and extract data presented directly in textual explanations of unfamiliar data; interpret text and diagrams describing a simple pattern; perform calculations including working with large numbers, calculations with speed and time, conversion of units (for example from an annual rate to a daily rate). They

Level	What students can typically do
	understand place value involving mixed 2- and 3-decimal values and including working with prices; and are typically able to order a small series of (4) decimal values; calculate percentages of up to 3-digit numbers; and apply calculation rules given in natural language.
2	At Level 2, students can typically interpret simple tables to identify and extract relevant quantitative information; interpret a simple quantitative model (such as a proportional relationship) and apply it using basic arithmetic calculations. They are able to identify the links between relevant textual information and tabular data to solve word problems; interpret and apply simple models involving quantitative relationships; identify the simple calculation required to solve a straight-forward problem; carry out simple calculations involving the basic arithmetic operations, as well as ordering 2- and 3-digit whole numbers and decimal numbers with one or two decimal places, and calculate percentages.
1a	At Level 1a, students are typically able to solve basic problems in which relevant information is explicitly presented, and the situation is straightforward and limited in scope. They are able to handle situations where the required computational activity is obvious and the mathematical task is basic, such as performing one or two simple arithmetic operations with whole numbers or percentages. Students at this level can manipulate quantitative information to make it amenable to computational analysis, such as determining the total number of points earned by teams given a record of their wins and losses.
1b	At Level 1b, students can solve straight-forward problems that require single arithmetic operations with whole numbers or retrieving numerical information from a table or chart. For example, students can total the columns of a simple table and compare the results, or they can read and interpret a simple table of monetary amounts or a work schedule to satisfy a situation with a single constraint.
1c	<i>There were no items in the PISA 2022 Mathematics assessment to describe this level on the scale.</i>

**Annex Table 17.A.10. Summary descriptions of the proficiency levels on the mathematical content subscale: Space and shape**

Level	What students can typically do
6	At Level 6, students are able to solve complex problems involving multiple representations or calculations; identify, extract, and link relevant information, for example by extracting relevant dimensions from a diagram or map and using scale to calculate an area or distance; they use spatial reasoning, significant insight and reflection, for example by interpreting text and related contextual material to formulate a useful geometric model and applying it taking into account contextual constraints; they are able to recall and apply relevant procedural knowledge from their mathematical knowledge base such as in circle geometry, trigonometry, Pythagoras's rule, or area and volume formulae to solve problems; and they are typically able to generalise results and findings, communicate solutions and provide justifications and argumentation.
5	At Level 5, students are typically able to solve problems that require appropriate assumptions to be made, or that involve reasoning from assumptions provided and taking into account explicitly stated constraints, for example in exploring and analysing the layout of a room and the furniture it contains. They solve problems using theorems or procedural knowledge such as symmetry properties, or similar triangle properties or formulas including those for calculating area, perimeter or volume of familiar shapes; they use well-developed spatial reasoning, argument and insight to infer relevant conclusions and to interpret and link different representations, for example to identify a direction or location on a map from textual information.
4	Students at Level 4 typically solve problems by using basic mathematical knowledge such as angle and side-length relationships in triangles, and doing so in a way that involves multistep, visual and spatial reasoning, and argumentation in unfamiliar contexts; they are able to link and integrate different representations, for example to analyse the structure of a three dimensional object based on two different perspectives of it; and typically they can compare objects using geometric properties.
3	At Level 3, students are able to solve problems that involve elementary visual and spatial reasoning in familiar contexts, such as calculating a distance or a direction from a map or a GPS device; they are typically able to link different representations of familiar objects or to appreciate properties of objects under some simple specified transformation; and at this level students can devise simple strategies and apply basic properties of triangles and circles, and can use appropriate supporting calculation techniques such as scale conversions needed to analyse distances on a map.
2	At Level 2, students are typically able to solve problems involving a single familiar geometric representation (for example, a diagram or other graphic) by comprehending and drawing conclusions in relation to clearly presented basic geometric properties and associated constraints. They can also evaluate and compare spatial characteristics of familiar objects in a situation where given constraints apply (such as comparing the height or circumference of two cylinders having the same surface area; or deciding whether a given shape can be dissected to produce another specified shape).
1a	Students at Level 1a can typically recognise and solve simple problems in a familiar context using pictures or drawings of familiar geometric objects and applying basic spatial skills such as recognising elementary symmetry properties, or comparing lengths or angle sizes, or using procedures such as dissection of shapes.
1b	<i>There were no items in the PISA 2022 Mathematics assessment to describe this level on the scale.</i>
1c	<i>There were no items in the PISA 2022 Mathematics assessment to describe this level on the scale.</i>

### Annex Table 17.A.11. Summary descriptions of the proficiency levels on the mathematical content subscale: Uncertainty and data

Level	What students can typically do
6	At Level 6, students are able to interpret, evaluate and critically reflect on a range of complex statistical or probabilistic data, information and situations to analyse problems. Students at this level bring insight and sustained reasoning across several problem elements; they understand the connections between data and the situations they represent and are able to make use of those connections to explore problem situations fully; they bring appropriate calculation techniques to bear to explore data or to solve probability problems; and they can produce and communicate conclusions, reasoning and explanations.
5	At Level 5, students are typically able to interpret and analyse a range of statistical or probabilistic data, information and situations to solve problems in complex contexts that require linking of different problem components. They can use proportional reasoning effectively to link sample data to the population they represent, can appropriately interpret data series over time and are systematic in their use and exploration of data. Students at this level can use statistical and probabilistic concepts and knowledge to reflect, draw inferences and produce and communicate results.
4	Students at Level 4 are typically able to activate and employ a range of data representations and statistical or probabilistic processes to interpret data, information and situations to solve problems. They can work effectively with constraints, such as statistical conditions that might apply in a sampling experiment, and they can interpret and actively translate between two related data representations (such as a graph and a data table). Students at this level can perform statistical and probabilistic reasoning to make contextual conclusions.
3	At Level 3, students are typically able to interpret and work with data and statistical information from a single representation that may include multiple data sources, such as a graph representing several variables, or from two simple related data representations such as a simple data table and graph. They are able to work with and interpret descriptive statistical, probabilistic concepts and conventions in contexts such as coin tossing or lotteries and make conclusions from data, such as calculating or using simple measures of centre and spread. Students at this level can perform basic statistical and probabilistic reasoning in simple contexts.
2	Students at Level 2 are typically able to identify, extract and comprehend statistical data presented in a simple and familiar form such as a simple table, a bar graph or pie chart; they can identify, understand and use basic descriptive statistical and probabilistic concepts in familiar contexts, such as tossing coins or rolling dice. At this level students can interpret data in simple representations, and apply suitable calculation procedures that connect given data to the problem context represented.
1a	At Level 1a, students can typically read and extract data from charts or two-way tables, and recognise how these data relate to the context. Students at this level can also use basic concepts of randomness to identify misconceptions in familiar experimental contexts, such as flipping a coin.
1b	Students at Level 1b, can typically read information presented in a well-labelled table to locate and extract specific data values while ignoring distracting information.
1c	<i>There were no items in the PISA 2022 Mathematics assessment to describe this level on the scale.</i>

### Annex Table 17.A.12. Cutpoints for the Reading Scale

Cut point	Level Name
698.32	Level 6
625.61	Level 5
552.89	Level 4
480.18	Level 3
407.47	Level 2
334.75	Level 1a
262.04	Level 1b
189.33	Level 1c

### Annex Table 17.A.13. Cutpoints for the Science Literacy Scale

Cut point	Level Name
707.93	Level 6
633.33	Level 5
558.73	Level 4
484.14	Level 3
409.54	Level 2
334.94	Level 1a

260.54	Level 1b <sup>1</sup>
--------	-----------------------

Note: 1. Level 1b bandwidth is slightly narrower than others.

### Annex Table 17.A.14. Cut points for the Financial Literacy Scale

Cut point	Level Name
624.63	Level 5
549.86	Level 4
475.10	Level 3
400.33	Level 2
325.57	Level 1

# 18 PISA 2022 Innovative Domain

## Test Design and Test Development

### Introduction

This chapter describes the assessment design framework for the PISA 2022 innovative domain of creative thinking as well as the processes used by the PISA Core B contractor, ACT, the PISA Secretariat, and the international test development team to develop the creative thinking assessment for the PISA 2022 cycle.

Activities undertaken in the context of the innovative domain test design and development included the following:

- The creation of a Creative Thinking Expert Group (CTEG) to guide the assessment framework, test design and test development;
- The development of a creative thinking assessment framework;
- The assessment design and development;
- A series of small-scale validation studies;
- Field trial activities; and
- The main survey administration.

### The role of the Creative Thinking Expert Group (CTEG) in the framework and item development

As the contractor for the creative thinking instrument development, Core B was responsible for working with the creative thinking expert group (CTEG) and the PISA Secretariat. Work focused on understanding the CTEG and PISA Secretariat's vision for the creative thinking assessment framework as well as the range and types of items to be developed for the test and questionnaire instruments. The PISA Secretariat and CTEG members began work on the framework in September 2017 finalised the framework in September 2022. Core B's work with the PISA Secretariat and CTEG began in February 2018 and focused on the following tasks:

- describing the kinds of items needed to assess the skills and abilities in each domain as 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 testing functionalities that would be desirable to develop for measuring the construct and that would be feasible to implement in the context of the PISA 2022 administration.

Work with the CTEG continued beyond the initial meeting in February 2018 through the entire phase of instrument development and during data analysis. CTEG members played an important role in reviewing

and providing feedback on the assessment tasks as they were developed, providing input into the analysis of the data from multiple small scale validation exercises and the field trial(s), approving the set of items for the main survey administration, and working with instrument development and data analysis staff to develop the described scales and performance level descriptors used for reporting the PISA 2022 creative thinking results.

## PISA 2022 creative thinking assessment framework

The PISA Secretariat, together with guidance from the CTEG, developed the PISA 2022 creative thinking assessment framework. The PISA 2022 creative thinking assessment focused on the creative thinking processes that can be reasonably expected from 15-year-old students around the world. It does not aim to single out exceptionally creative individuals but rather to describe the extent to which students can think creatively when searching for and expressing ideas, and to describe how this capacity is related to teaching approaches, school activities and other features of education systems.

The main objective of PISA is to provide internationally comparable data on students' competencies that have clear implications for education policies and pedagogies. In the context of the PISA 2022 assessment, the creative thinking processes in question therefore need to be malleable through education; the different enablers of these thinking processes in the classroom context need to be clearly identified and related to performance in the assessment; the content domains covered in the assessment need to be closely related to subjects taught in common compulsory schooling; and the test tasks should resemble real activities in which students engage, both inside and outside of their classroom, so that the test has some predictive validity of creative achievement and progress in school and beyond.

While closely related to the broader construct of creativity, the PISA 2022 assessment focuses on creative thinking understood as the cognitive processes that are required to engage in creative work. Creative thinking was considered a more appropriate construct to assess in the context of PISA as it is a malleable individual capacity that can be developed through practice, and it refers more to specific cognitive processes than to the subjective quality of an output.

PISA defines creative thinking as:

*The competence to engage productively in the generation, evaluation, and improvement of ideas, that can result in original and effective solutions, advances in knowledge, and impactful expressions of imagination.*  
(OECD, 2023<sup>[1]</sup>)

The PISA definition builds on definitions of creativity and creative thinking found in the literature, following a comprehensive review, and it was developed with the guidance of a wider interdisciplinary group of experts in the field (the CTEG). The definition is aligned with the cognitive processes and outcomes associated with “little-c” creativity – in other words, it reflects the types of creative thinking that 15-year-old students around the world can reasonably demonstrate in everyday contexts. It emphasises that students need to learn to engage productively in generating ideas, reflecting upon ideas by valuing their relevance and novelty, and iterating upon ideas before reaching a satisfactory outcome. This definition of creative thinking applies to learning contexts that require imagination and the expression of one's inner world, such as creative writing or the arts, as well as contexts in which generating ideas is functional to the investigation of problems or phenomena.

### ***The competency model***

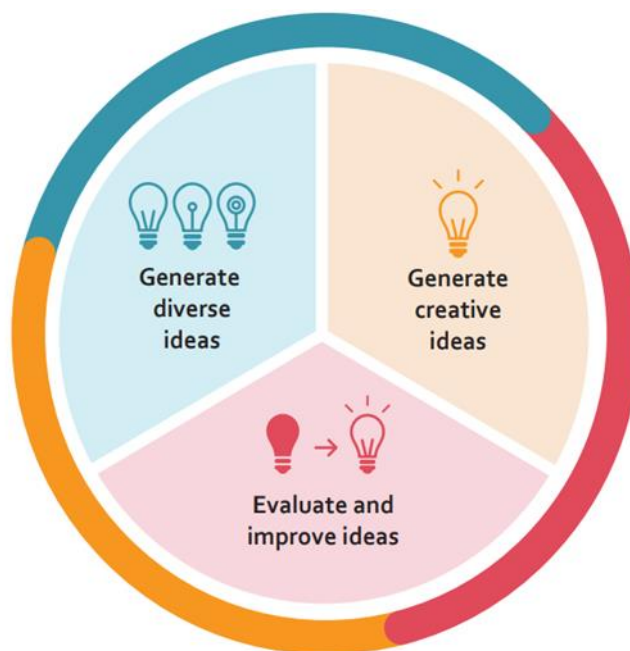
Three cognitive facets support creative thinking and constitute the competency model for the PISA 2022 creative thinking assessment (see Figure 18.1). These three facets are:

- generate diverse ideas;

- generate creative ideas; and
- evaluate and improve ideas.

These three facets reflect the PISA definition of creative thinking and incorporate both divergent cognitive processes (the ability to generate diverse ideas and to generate creative ideas) and convergent cognitive processes (the ability to evaluate other people's ideas and identify improvements to those ideas). “Ideas” in the context of the PISA assessment can take many forms, and the test units provide a meaningful context and sufficiently open tasks in which students can demonstrate their capacity to produce different ideas and think outside of the box.

**Figure 18.1. The PISA 2022 competency model for creative thinking**



### *Generate diverse ideas*

Typically, attempts to measure creative thinking have focused on the number of ideas that individuals are able to generate – often referred to as ideational fluency. Going one step further is ideational flexibility, or the capacity to generate ideas that are different to each other. When it comes to measuring the quality of ideas that an individual generates, some researchers have argued that fundamentally different ideas should be weighted more than similar ideas (Guilford, 1956<sup>[2]</sup>). The facet ‘generate diverse ideas’ of the competency model encompasses these notions and refers to a student’s capacity to think flexibly by generating multiple distinct ideas. Test items for this facet present students with a stimulus and ask them to generate two or three appropriate ideas in response that are as different as possible from one another.

### *Generate creative ideas*

The literature generally agrees that creative ideas and outputs are defined as being both novel and useful (Plucker, Beghetto and Dow, 2004<sup>[3]</sup>). Expecting 15-year-olds around the world to generate ideas that are completely unique or novel is clearly neither a feasible nor appropriate approach for the PISA assessment. Instead, originality represents a useful concept as a proxy for measuring the novelty of ideas. Defined by Guilford (1950<sup>[4]</sup>) as “statistical infrequency”, originality encompasses the qualities of newness, remoteness, novelty or unusualness, and generally refers to deviance from patterns that are observed



within the population at hand. In the PISA assessment context, originality is therefore a relative measure established with respect to the responses of other students who complete the same task.

The facet ‘generate creative ideas’ focuses on a student’s capacity to generate appropriate and original ideas. This dual criterion ensures the measurement of creative ideas – ideas that are both original *and* of use – rather than ideas that make random associations that are original yet not meaningful. Test items for this facet present students with a stimulus and ask them to develop one original idea in response.

### *Evaluate and improve ideas*

Evaluative cognitive processes help to identify and remediate deficiencies in initial ideas as well as ensure that ideas or solutions are appropriate, adequate, efficient and effective (Cropley, 2006<sup>[5]</sup>). They often lead to further iterations of idea generation or the reshaping of initial ideas to improve a creative outcome. Evaluation and iteration are thus at the heart of the creative thinking process. The facet ‘evaluate and improve ideas’ focuses on a student’s capacity to evaluate limitations in ideas and improve their originality. To reduce problems of dependency across items in the test, students are not asked to iterate upon their own ideas but rather to modify a provided “idea”. Test items for this facet thus present students with a given scenario and idea and ask them to suggest an original improvement in response, defined as a change that preserves the essence of the initial idea but that adds or incorporates original elements.

### **Task contexts: domains of creative thinking**

The literature suggests that the larger the number of domains included in an assessment of creative thinking, the better the coverage of the construct given that creative thinking draws on both domain-general and domain-specific resources. The choice of which domains to include in the PISA test was thus a central design question. Given the age and diversity of PISA test takers (15 years-old in over 60 countries), and the fact that domain knowledge is an important enabler of creative thinking, the domain contexts included in the assessment needed to be familiar and accessible to most students around the world, be relevant to schooling, reflect realistic manifestations of creative thinking that 15-year-olds could achieve in a constrained test context, and represent a sufficiently diverse coverage of different types of “everyday” creative thinking as reflected in the literature. Further practical constraints, including the available testing time (a maximum of one hour for the creative thinking test) and testing technology, also informed design choices.

Taking these main constraints into account, the PISA test of creative thinking includes tasks situated within four distinct domain contexts:

- written expression;
- visual expression;
- social problem solving; and
- scientific problem solving.

In the PISA test, the written and visual expression domains involve communicating one’s imagination to others, and creative work in these domains tends to be characterised by originality, aesthetics, imagination, and affective intent and impact. In contrast, the social and scientific problem-solving domains involve investigating and solving open problems. They draw on a more functional employment of creative thinking that is a means to a better end, and creative work in these domains is characterised by ideas or solutions that are original, innovative, effective and efficient.

The inclusion of tasks situated in several domain contexts will allow the PISA 2022 creative thinking test to provide information about students’ strengths and weaknesses in creative thinking across countries. The test items were distributed across the three facets and four domain contexts to allow for a range of

opportunities for students to engage and express creative thinking. Annex Table 18.A.2 sets out the distribution of items across facets and domains for the field trial(s) and the main survey administration.

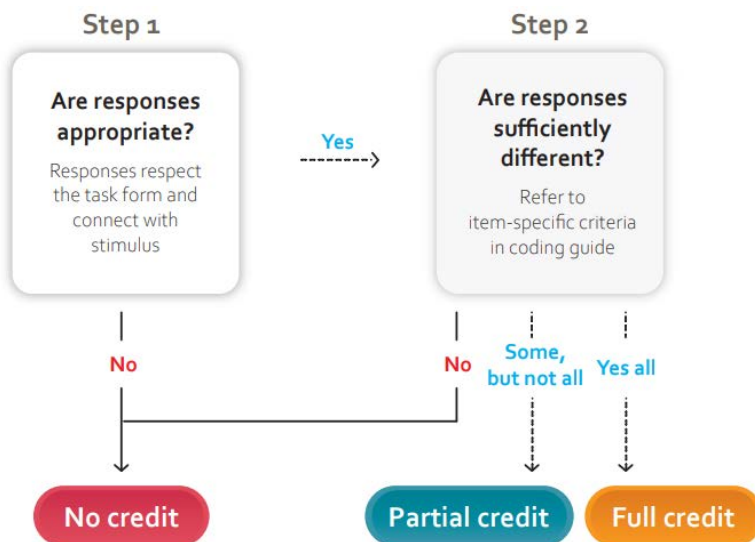
### Coding and scoring processes

Every task in the PISA creative thinking test is open-ended and scoring student responses relies on human judgement following detailed scoring rubrics and well-defined coding procedures. All items corresponding to the same facet of the competency model (i.e. 'generate diverse ideas', 'generate creative ideas' and 'evaluate and improve ideas') apply the same general coding procedure. However, as the form of response varies by domain and task (e.g. a title, a solution, a design, etc.), so do the item-specific criteria for evaluating whether an idea is different or original. ACT developed detailed coding guides to describe the item-specific criteria for each item and provide annotated example responses to help human coders score consistently.

#### Scoring of 'generate diverse ideas' items

All items corresponding to the 'generate diverse ideas' facet of the competency model require students to provide two or three responses. The general coding procedure for these items involves two steps, as summarised in Figure 18.2. First, coders must determine whether responses are appropriate. Appropriate in the context of the creative thinking assessment means that students' responses respect the required form and connect (explicitly or implicitly) to the task stimulus. Second, coders must determine whether responses are sufficiently different from one another based on item-specific criteria described in the coding guide.

Figure 18.2. General coding process for 'generate diverse ideas' items



The item-specific criteria are as objective and inclusive as possible of the range of different potential responses. For example, for a written expression item, sufficiently different ideas must use words that convey a different meaning (i.e. are not synonyms). For items in the problem-solving domains, the coding guides list pre-defined response categories to help coders distinguish between similar and different ideas. The coding guides provide detailed example responses and explanations for how to code each example.

Full credit is assigned where all the responses required in the task are both appropriate and different from each other. Partial credit is assigned in tasks requiring students to provide three responses and where two

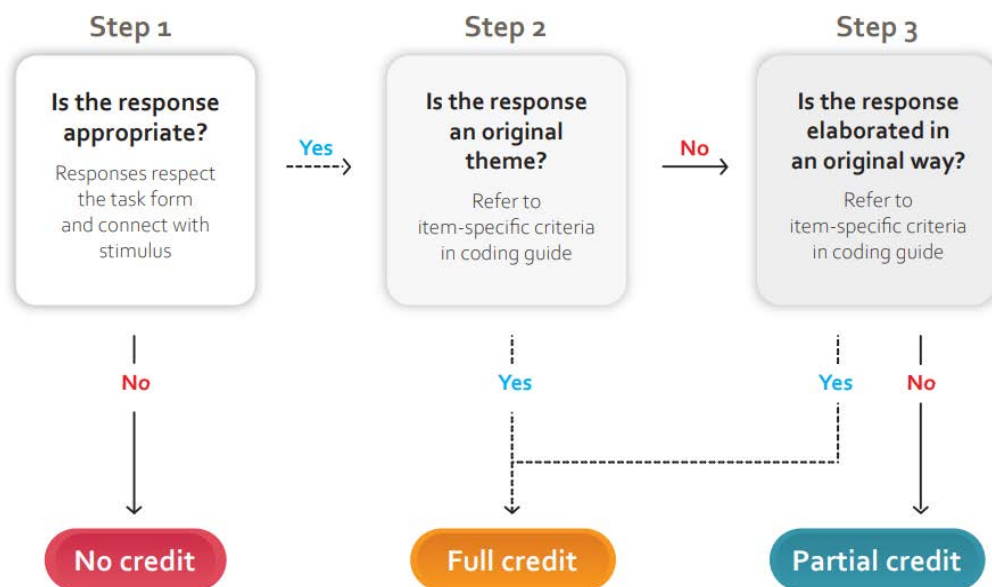
or three responses are appropriate, but only two are different from each other. No credit is assigned in all other cases.

### Scoring of 'generate creative ideas' items

All items corresponding to the facet 'generate creative ideas' of the competency model require a single response. The general coding procedure for these items involves two or three steps depending on the content of the response. First, as with all items, coders must determine whether the response is appropriate. Then, coders must determine whether the response is original by considering two criteria (see Figure 18.3).

An original idea is defined as a relatively uncommon idea with respect to the entire pool of student responses. The coding guide identifies conventional themes for each item according to the patterns of genuine student responses revealed in the validation studies. If a response does not correspond to a conventional theme as described in the coding guide, it is directly coded as original; however, if an idea does correspond to a conventional theme, then coders must determine whether it is original based on its elaboration. The coding guide provides item-specific explanations and examples of original ways to elaborate on conventional themes. For example, a student might add an unexpected twist to a story idea that otherwise centres on a conventional theme.

Figure 18.3. General coding process for 'generate creative ideas' and 'evaluate and improve' items



This twofold originality criteria ensures that the scoring model takes into account both the general idea and the details of a response. While this approach does not single out the most original responses in the entire response pool, it does ensure that the coding process is less susceptible to culturally-sensitive grading styles that favour middle points or extremes and it provides some mitigation against potential cultural bias in the identification of conventional themes across countries.

Full credit is assigned where the response is both appropriate and original. Partial credit is assigned where the response is appropriate only, and no credit is assigned in all other cases.

### *Scoring of 'evaluate and improve ideas' items*

All items corresponding to the facet 'evaluate and improve ideas' of the competency model require a single response and generally ask students to adapt a given idea in an original way rather than coming up with an idea from scratch. The general coding procedure for these items involves the same steps as those for the 'generate creative ideas' items. However, appropriate responses for these items must be both relevant and constitute an improvement. The threshold for achieving the appropriateness criteria for these items is thus somewhat strengthened with respect to items measuring the other two facets, as responses must explicitly connect to the task stimulus and attempt to address its deficiencies. The coding guide provides item-specific criteria, examples and explanations to help orient coders. For responses considered appropriate, coders must then establish the originality of the improvement by considering the same two originality criteria as for 'generate creative ideas' items.

Full credit is assigned where the response is both appropriate and an original improvement. Partial credit is assigned where the response is appropriate only, and no credit is assigned in all other cases.

## **PISA 2022 innovative domain test assembly design**

According to the PISA assessment design, about 28% of the sample of PISA students were administered the creative thinking assessment. Students who took the creative thinking assessment spent one hour on creative thinking test items with the remaining hour of testing time assigned to one of the other core domains (mathematics, reading or scientific literacy).

The creative thinking items were organised into test units. The units vary in terms of the facets that are measured (i.e. generate diverse ideas, generate creative ideas, and evaluate and improve ideas), the domain context (i.e. written expression, visual expression, social problem solving, or scientific problem solving) and the duration of the unit (guidelines of between 5 and 15 minutes). Some units are composed of a single item and some units have multiple items. Dependencies between items within units was minimised.

The creative thinking units were then organised into five, mutually exclusive 30-minute blocks or clusters. The clusters were rotated according to the integrated design presented in Chapter 3 of this Technical Report.

Constructed-response tasks accounted for 92% of the items in the creative thinking assessment. The tasks typically call for a written response, ranging from a few words (e.g. cartoon caption or scientific hypothesis) to a short text (e.g. creative ending to a story or explanation of a design idea). Some constructed-response items call for a visual design response (e.g. designing a poster combining a set of given shapes and stamps) that is supported by a simple drawing editor tool. The assessment also included 2 items that were part of an interactive simulation-based task and two (possible) multiple-choice items where students are given the option to select a previously suggested idea or to generate a new idea.

## **PISA 2022 innovative domain assessment design and development**

Test development for the PISA 2022 creative thinking assessment cycle began in early-2018 and focused on the development of items for a computer-based assessment. Through a process that included both CTEG contributions, as well as country submission and country review, Core B along with the PISA Secretariat selected an initial set of unit and item scenarios. Core B test developers then further developed the unit and item scenarios. The PISA Secretariat reviewed all unit scenarios and items early in the review process, prior to country reviews, to ensure the items fulfilled the goals of the assessment framework.

The developed units were submitted for translatability review at the same time that they were released for country review. 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 2022 cycle alerted test developers to both general wording patterns and specific item wording that are known to be problematic for some translations and suggested alternatives. This allowed test 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.

To ensure that the creative thinking assessment items were understood the same way across linguistic and cultural groups, participating countries also engaged in several cycles of review of the test material to help identify items that may be likely to suffer from cross-cultural bias. This enabled problematic cultural and linguistic characteristics to be identified during the early stages of the assessment development process. Countries had two weeks to perform reviews and submit feedback on all draft stimuli and items.

Preparation of the French source version for all of the test units provided another opportunity to identify issues with the English source version related to content and expression. The development of the two source versions helped to identify instances where wording may prove problematic for translation and 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.

### ***Cognitive laboratories***

Experienced testing professionals were engaged to conduct cognitive laboratory exercises with students in Australia, Singapore and the United States. A total of 66 students across the three countries participated in these cognitive labs in the period between August and September 2018 on set of eight prototype units across the four domain contexts. In the format of concurrent and retrospective thinking-out-loud exercises, students around the age of the PISA population were asked to explain their thought processes while completing the test items and to point out any difficulties or misunderstandings in the instructions or stimulus material. After students completed all of the tasks within each unit, they were asked to answer a series of probing questions about their experience working through the tasks including specific questions on the comprehensibility of the task prompt and perceived difficulty of the task. Students also went through each task a second time to verbalise any thoughts they had had when working through the task. The cognitive laboratories helped to evaluate whether students could understand what they were asked to do during the test, whether students perceived the tasks as engaging, excessively demanding or frustrating, and whether they needed more clarifications to be added to the task prompts.

The analysis of the information collected during these sessions, as well as from video recordings, identified opportunities for the revision and optimisation of items as well as to correct several identified bugs in the testing platform (ACT, 2018<sup>[6]</sup>). Insights from the cognitive laboratories included:

- **Refining the number of required responses in ‘generate diverse ideas’ items.** In the prototype units tested in the cognitive labs, students could enter as many responses as they wished on these items. In general, students created up to three responses for these items with relative ease but expended considerable effort to move beyond three responses. Moreover, their fourth or fifth responses were rarely their most creative ones. As a result, in successive revisions of the test material, the test development team decided to ask students for up to three responses only and emphasise in the task prompt that students should aim to provide responses that are as different as possible from each other.
- **Choosing the number of required responses in time-intensive items.** Some of the prototype items required a significantly greater investment of time, elaboration and careful execution than others. It was evident from students’ feedback and actual responses to some tasks that asking them to iterate upon or produce more than one response could easily generate fatigue; it was thus

decided that in these types of time-intensive tasks, students should be asked to generate no more than one response.

- **Providing guidance about time required on the task.** In the cognitive labs, students could spend as much time as they wished on each task, although many expressed the need to have some kind of guidance on timing. Different solutions for providing guidance on time usage were considered; the test development team decided to provide an indication of the maximum amount of time that students should spend on each task in the respective task prompt to help students manage their time.
- **Clarifying task expectations and instructions.** Some students requested further clarification on certain terms used in the tasks prompts (e.g. “original”). The prompts were subsequently revised to reduce subjective interpretations of such terms as much as possible. For example, a clarification was added to explain that “original” refers to a solution that other students might not have thought of (clearly associating originality to statistical frequency).
- **Selecting the right images as task stimuli.** Several tasks use a visual stimulus and ask students to engage in idea association in order to generate a response inspired by the image. Some images used in the prototype units evoked associations that were strongly culturally-mediated (for example, some students thought of the Beatles’ song when they saw the image of a yellow submarine). While cultural influences upon student responses cannot be completely eliminated, the development team revised any images that were clearly susceptible to inspiring culture-specific associations.
- **Defining features of the drawing tool.** Most students rapidly understood how to use the drawing tool provided in the platform. However, in some cases students clearly lost precious time trying to complete specific actions that were not immediately intuitive (e.g. deleting an object.) These issues have been addressed by including a tutorial on the use of the drawing tool. The test development team evaluated the potential advantages and disadvantages of including additional features in the drawing tool and decided to keep a relatively simple tool with limited graphical instruments to limit any potential unfair advantage to those students who are more proficient in doing graphical work on a computer. The analysis of responses from the cognitive labs confirmed that it is possible to generate highly creative outputs using only a limited version of the drawing tool.

Six of the eight prototype units were further developed after the cognitive labs for inclusion in subsequent validation studies, while two units were abandoned at this stage due to unsatisfactory performance in the cognitive labs. A further set of units were also developed in accordance with the insights from the cognitive laboratories.

### ***Small scale validation exercises***

Further small-scale validation exercises were conducted in parallel to the overall test development process, in an iterative manner, to observe how the then-current test materials functioned under similar test conditions to the field trial and main survey. The purpose of these validation studies was several-fold:

- to provide evidence on the performance of the creative thinking assessment in PISA-like classroom settings;
- to collect sample student responses in multiple countries to inform the development of the coding and scoring guides;
- to assess the inter-rater reliability of human coded items (i.e. the agreement between raters);
- to gain insights into the difficulty of the items;
- to determine the extent to which a creative thinking score or sub-scores could be obtained from the creative thinking assessment; and

- to gain preliminary insights on the essential coder training materials and processes needed for human coders.

A total of 703 15-year-old students from Singapore (n=206), Australia (n=234) and Canada (n=263) participated in the first validation study between October to November 2018. Samples were recruited through the PISA National Project Managers and coordinated with the PISA Secretariat. The validation study instrument included 12 fully functional prototype units delivered in 3 test forms, with 4 units per form. Each form contained one unit per domain.

The coding of the units was carried out according to the preliminary coding guides developed by ACT. Student responses were scored by a team of professional scorers at ACT. As a group, the team reviewed and assigned scores to 5% of the available responses for each task, which enabled scorers to build a common understanding of the coding procedures. Each response was then coded independently by two scorers. Any questions or issues that arose during the scoring of the data were referred to the Scoring Supervisor and the Assessment Design team at ACT.

An analysis of the genuine student data indicated items that did not perform as intended and informed evidence-based improvements to the test material, as well as development of and improvements to coder training material such as the coding guide (ACT, 2019<sup>[7]</sup>). The validation study also helped to refine the methodology followed for scoring students' responses – in particular, it informed the introduction of a double criteria for coding the originality of responses taking into consideration both the originality of the theme of a students' response and the originality of their approach – and provided genuine responses for the international coder workshops.

A total of 202 15-year-old students from the Republic of South Africa participated in the second validation study from February to March 2019. ACT and the PISA Secretariat partnered with the Care for Education in Republic of South Africa, with support from the LEGO Foundation, to carry out the validation study. It included 16 units, delivered in two test forms. This validation study was delivered on paper to simplify administration procedures (only a limited time was available for the recruitment of schools and it was not possible to condition participation to the availability of computer equipment). Each test form contained the same units, but the order of the units presented to students varied to mitigate potential order effects on performance.

In the second validation study, student responses were scored by a team of non-professional scorers at Care for Education that were trained by the scoring team at ACT following a standard training process and with the support of the international coding guides. Similarly to the first validation study, each response was coded independently by two scorers.

The second validation study provided valuable insights into the success of revisions to the units throughout the test development process. In general, the performance of the creative thinking item pool in the second validation study improved upon the performance of the items included in the first validation study, particularly in terms of inter-rater reliability.

## Field trial

The field trial for creative thinking was initially scheduled for 2020; however, this timeline was disrupted by the COVID-19 global pandemic meaning only a limited field trial (LFT) was carried out, with findings further investigated during a second administration of the field trial in 2021. The LFT conducted in 2020 with 11 countries provided preliminary evidence in support of:

1. the psychometric quality of the PISA 2022 creative thinking assessment units in terms of their validity, reliability, and comparability across participating countries;
2. the ability to construct a creative thinking scale and, possibly, subscales;

3. the inclusion of all the creative thinking units and forms in Field Trial 2021; and
4. further enrichment of the coder training materials utilised in coder training for the full field trial in 2021 and the main survey administration in 2022 (ACT, 2020<sup>[8]</sup>).

In 2021, a further field trial was conducted with 44 countries to provide additional evidence of the validity and reliability of the creative thinking assessment.

### ***Field trial coder training***

Among the total 38 items administered, two items were machine-scored (the simulation-based items) and the remaining 36 items were human-scored items. For the human-scored items, all coding processes were performed by each country's coders. The ACT team provided international coder training and supported the national coding teams through a standard PISA query service.

#### *Limited field trial (2020)*

The coding guide for the PISA 2022 creative thinking assessment was developed by test developers and performance scoring experts at ACT, with the support of the PISA Secretariat. Coder training procedures and materials were informed by the cognitive labs and validation studies and included examples of genuine student responses.

The English master version of the coding guide was released in a draft form prior to the in-person PISA International Coder Training meeting in January 2020. The training objectives included developing a foundational understanding of the creative thinking construct and an in-depth understanding of the coding processes so that attending representatives would be prepared to train coders in their countries using the provided materials. Test developers and performance scoring experts from ACT, with the support of the PISA Secretariat, facilitated discussions at that meeting. The coding guide used in the limited field trial was finalised based on these discussions. The updated English version of the coding guide and the French source version were subsequently released to countries in February 2020 prior to the beginning of the limited field trial data collection period.

#### *Field trial (2021)*

The International Coder Training meeting for creative thinking ahead of the full field trial was held virtually over 5 days due to the COVID-19 pandemic in February 2021. Performance scoring experts from ACT developed online coding training modules and facilitated an interactive coder training workshop, held with representatives from the participating countries in the 2021 field trial prior to coding. To facilitate the online coder training, ACT's team developed comprehensive exemplar sets consisting primarily of authentic student responses that were selected and intended to demonstrate a typical response for each credit level and theme assignment (i.e. codes 00, 11, 12, 13, 21, 22, 23, etc., with code 29 used to designate an unlisted theme). Discussion was also dedicated to reinforcing understanding and consensus about the coding rules for each item to better ensure consistency of coding within and between countries.

Facilitators reviewed the layout of the coding guide, general coding principles, common problems, and guidelines for applying special codes. Workshop materials were optimised based on feedback from the LFT coder training, LFT coder queries and translation referee updates to the earlier version of the coding guide. Attendees were required to code the workshop materials (i.e. the exemplar sets) "live" during the interactive workshop; where there were disagreements about the coding for an item, those were discussed in detail so that all attendees understood, and would be able to follow, the intent of the coding guides. In some instances, disagreements – particularly those highlighting possible cultural bias – led to modifications of the coding guide and/or workshop materials.



### ***Preparation of the field trial data collection instruments***

The process for creating the field trial national student delivery system (SDS) began with the assembly and testing of the master SDS, followed by the process for assembling national versions of the field trial SDS. After all components of the national materials were locked, including the questionnaires and cognitive instruments, the student delivery system was assembled and tested first 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 final systems were released for the field trial. The PISA 2022 creative thinking assessment was only administered on computers.

### ***Field trial coding procedures***

The field trial design required that two independent coders review and code each student's responses at a credit level of either 0,1 (i.e. no credit or credit), or 0, 1, or 2 (i.e. no credit, partial credit or full credit), thus generating inter-rater reliability at the credit level. In addition, two selected English-fluent bilingual coders from each country reviewed and coded 30 pre-designated anchor responses to verify coder reliability across countries. These anchor responses were selected from earlier validation studies conducted in Australia, Canada, Colombia, Singapore and South Africa, and represented a range of responses at all credit levels (ACT, 2019<sup>[7]</sup>). Inter-rater reliability (IRR) on the anchor responses across all items and coder pairs was high (0.71). The average quadratic Kappa was also high (0.79).

For the items measuring either the 'generate creative ideas' or 'evaluate and improve ideas' facets, coders were required to use a second digit to indicate the primary theme of each response that earned either partial or full credit. Partial credit responses could only be coded using values of 1-3 as their second digit (i.e. codes 11, 12 or 13), to represent correspondence with the initial conventional themes designated in the coding guide based on an analysis of available student responses in the validation studies; however, responses that received full credit could use up to 9 different values for the second digit (i.e. codes 21 through 29), with the ninth value representing all themes not associated with themes 1-8. The resulting data informed distinctions between "conventionality" and "unconventionality" of themes across a diverse international student cohort.

### ***Field trial coder queries procedures***

As was the case during previous cycles, Core A set up and maintained a coder query service for the 2020 and 2021 field trials. Countries were encouraged to send coder queries to the service so that a common adjudication process was consistently applied to all coder questions about constructed-response items. Core B test developers and performance scoring experts from ACT reviewed and responded to coder queries that were specific to the creative thinking test.

In addition to responses to new queries, Core B curated a selection of queries to include in the Coder Query Log containing accumulated responses from previous cycles of PISA. This helped foster consistent coding of creative thinking items. The query log was regularly updated and posted for National Centres on the PISA portal as new queries were received and processed.

### ***National item review post-Field Trial***

The item feedback process began in August 2021 and concluded in October 2021 and was conducted in two phases. Phase 1 occurred before countries received their field trial data and Phase 2 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 turnaround 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 B. Following the 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 then reviewed by national teams. As was the case in Phase 1, countries were asked to provide comments about these specific items in instances where they could identify serious errors. Requests for corrections were reviewed by Core B and, where approved, implemented.

### ***Field Trial outcomes***

The 2021 Field Trial data analyses addressed the issue of construct and score validity and reliability, within and across countries, in addition to differential item functioning. Following the field trial data collection, the items were analysed for inter-rater reliability on anchor responses, inter-rater reliability on all responses, average Quadratic Kappa, item category response functions, item quality, and item omit and not-reached rates. Items that exceeded the omit and not-reached rates were identified and investigated; in some cases, this could be attributed to technical issues with some items during the administration of the test, and cluster placement was also considered to be a contributing factor.

Other analyses of the data included item difficulty, item discrimination, item response time, position effects, IRT scaling, item model fit, IRT parameters and student theta estimates, the evaluation of sub-scores on domain and facet levels, and differential item functioning (DIF) via the item-total score curves from different country-by-language groups. Any flagged items for DIF were further reviewed in terms of their sample size, contents, translations and coding guides (i.e. verified translation vs non-verified translation of coding guides), student responses (indications of misunderstanding), performance in alternative languages for that country, performance on similar items in assessment for that country/language, performance on the other items in that unit, additional item flags for that item, LFT data vs. FT data, and planned optimisations for that item (e.g., theme changes, coding optimisations or cluster placement).

Due to the operational timeline in PISA, it was not possible to include new items in the creative thinking test after this phase and no substantial modifications were made to existing test items, i.e. poorly performing items were removed from the test item pool to ensure a proper coverage of the construct in the main survey. Following the field trial analyses, one unit consisting of two items was removed (see Annex Table 18.A.2).

In summary, the findings from the field trial analysis supported:

1. the psychometric quality of the PISA 2022 creative thinking assessment units in terms of their validity, reliability, and comparability across participating countries;
2. the ability to construct a creative thinking scale; and
3. the inclusion of 20 of the 21 creative thinking units administered in the field trial for administration in the 2022 main survey.

The field trial(s) also generated insights for the further enrichment of the coder training materials, including the coding guide, prior to the 2022 main survey. Substantial work was undertaken including reviewing large amounts of genuine student responses, conducting an additional frequency analysis of response themes, and identifying instructions that caused coding issues by being absent, too vague or too restrictive. This resulted in substantial modifications of the coding guide, including updates to the designation of conventional and unconventional themes, the refinement of theme descriptions, the increased representation of exemplar responses, and edits to the item-specific instructions to facilitate effective and consistent coding.

## PISA 2022 main survey

The PISA 2022 main survey was conducted between March and December 2022. The majority of countries completed the main survey data collection by August 2022. 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 B contractors worked with countries to resolve any remaining issues and prepare the national instruments for the main survey.

### ***Item review and selection***

The PISA 2022 field trial provided evidence in support of the psychometric quality of the PISA 2022 creative thinking assessment units in terms of validity, reliability, and comparability across participating countries. Maintaining the same range of contexts from the field trial to the main survey provided good continuity and kept a consistent representation of skills and domains. Clusters were created following the final item selection and balanced based on the coverage of cognitive processes, the discrimination and difficulty of the items, and the total number of units and items. The duration of each unit was between 5 and 15 minutes. The units were organised into five mutually exclusive 30-minute blocks or clusters, and the clusters were rotated according to the integrated design presented in Chapter 3 of this Technical Report. The assessment aimed to achieve a good balance between units that situate creative thinking within the two thematic content areas (creative expression, and knowledge creation and problem solving) and the four domains.

The CTEG reviewed the field trial data and outcomes, the approach to item selection, the content and balance of the proposed main survey clusters, and signed off on the selection.

### ***Main survey coder training***

The main survey International Coder Training for creative thinking was held in February 2022. Analysis of student responses and coder queries during the field trial administration helped performance scoring experts from ACT improve upon the online coding training modules and other coder training and workshop materials. Additional sample responses were included in the coding guide to better illustrate different types of student responses. Workshop materials were also enhanced to include additional authentic student responses that better illustrated the boundaries between full credit, partial credit (where appropriate) and no credit.

The main survey coder training process was similar to that ahead of the 2021 field trial in that self-guided online training modules were completed before full-group discussions. The training objectives again included developing a foundational understanding of the construct and an in-depth understanding of the coding processes so that attending representatives would be prepared to train coders in their countries using the provided materials. Facilitators again reviewed the layout of the coding guide, general coding principles, common problems, guidelines for applying special codes, and workshop materials for each item. Following the international coder training, additional and final revisions were made to the coding guide in response to discussions that took place at the meeting.

### ***Preparation of data collection instruments***

The process for creating the main survey national student delivery system (SDS) followed the approach used during the field trial, beginning with the 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 first 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 final systems

were released for the main study. The PISA 2022 creative thinking assessment was only administered on computers.

### **Main survey coder queries**

The coder query service was again used in the main survey as it was in the field trial to assist countries in clarifying any uncertainty around the coding process or students' responses. Queries were reviewed and responses were provided by domain-specific teams including test developers and coding experts. Core B test developers and performance scoring experts from ACT reviewed and responded to queries specific to the creative thinking test. Relevant queries were included in the Coder Query Log, a resource maintained by Core A and accessible by all participant NPMs in the PISA Portal.

## **Data adjudication and approach to scaling the data for reporting**

In June 2023, Core A presented the Technical Advisory Group (TAG) with the PISA 2022 creative thinking data and preliminary psychometric analyses for data adjudication. Following the initial feedback of the TAG on the scalability of the data given the relatively low inter-item correlations and the creation of plausible values, the PISA Secretariat conducted further analyses of the creative thinking data including modifying some of the scoring rules with the goal of increasing the validity of inferences drawn from the creative thinking data, and improving the scalability and comparability across countries.

Following a thorough review of the data, the following changes were implemented:

- **Four items were dropped from the scaling.** The four items identified for exclusion were drawn from two units (one visual expression, and one scientific problem solving) and were all in the same test cluster. These four items showed poor discrimination and high omit rates, likely due to their position within the cluster.
- **The scoring rules for 14 items were modified.** All 'generate creative ideas' and 'evaluate and improve' items were reviewed following the main survey in terms of the distribution of double-digit codes across countries. The scoring process for these items required coders to use a second digit to indicate the primary theme of each response, and those coded using values of 1-3 as their second digit (i.e. 11, 21, 12, 22, 31 or 32) represented correspondence with the initial conventional themes designated in the coding guide. The double-digit codes were intended to serve as a mechanism through which to review the distribution of codes across countries and adjust the themes designated as conventional following the field trial and main survey. The number of conventional themes were modified for 14 of the 18 items corresponding to 'generate creative ideas' and 'evaluate and improve ideas' based on the results of the main survey to improve the validity of the scoring rules for these items and to align the scoring with the framework (i.e. originality as statistical infrequency, with respect to the responses of other students who completed the same task).
- **Responses submitted in fewer than 15 seconds were invalidated (i.e. converted to missing responses).** For most items in the creative thinking test, students must generate a written or visual artefact in response to a written or visual stimulus (i.e. task prompt with instructions and material for inspiration). The construct of creative thinking also aims to measure the cognitive processes associated with idea generation, evaluation and improvement, which are considered to be slow and thoughtful processes rather than reflective of opportunistic or rapid processes. For most items in the test, responses submitted within 15 seconds of viewing the item cannot be considered reflective of creative thinking processes. A review of the timing data for the items also showed a clear bimodal distribution of response submission, with one peak prior to 15 seconds and another peak a significant time afterwards. This modification was applied to all items, with the exception of

three: in two cases, students were able to select a response to a previous question akin to a multiple-choice mechanism; and in the other item, students were asked to generate a very short written artefact. In these three cases, it was judged that students could submit a response that reflected creative thinking processes within 15 seconds and thus no minimum response time was imposed.

In October 2023, the PISA Secretariat, Core A and the TAG reconvened for the data adjudication of the creative thinking data following the further analyses conducted by the PISA Secretariat and to finalise the reporting approach. The TAG recommended to report the creative thinking data according to a non-linear transformation of the “theta” scale, using the test-characteristic curve for a hypothetical test using the final pool of 32 creative thinking items and based on international item parameters. The advantages of this approach include:

- Reporting student performance according to a bounded scale (between 0-60, reflecting the maximum sum-score of all items) that is the same for all countries. This solution maintains the possibility to report performance on a scale, but signals a clear difference to the PISA scales used for the other domains and the broader “grain” size of the creative thinking scale signals its relative lower reliability compared to the other PISA scales (a 1 point change in the creative thinking scale reflects about 10% of a standard deviation).
- Scores can be easily interpreted in terms of the number of items correct on this specific test (rather than a more general reflection of students’ creative thinking ability applied to other performance tests), drawing attention to the actual test content and the framework that guided its development and facilitating the interpretation of the relatively high frequency of low scores in this test (i.e. students scored 0 on the test, rather than not having any creative thinking skill).
- Test scores differ more where the test has more information about students.
- The international database still includes 10 “plausible scores” per student.

### ***Performance level descriptors***

Following the data adjudication process and the finalisation of the scale for reporting the creative thinking data, the PISA Secretariat, in collaboration with Core B and the CTEG, defined performance level descriptors. Performance on the creative thinking scale was split into 6 performance levels.

#### *Level 1*

At level 1, students can generate very simple visual designs using isolated shapes or existing visual elements, and in some cases very short written artefacts (e.g. a few words), that require them to engage their imagination. In general, students at this level rely on obvious themes or idea associations as the basis for their response and struggle to generate more than one appropriate idea even for open and simple imagination tasks. These students typically generate simple visual or written artefacts with few details that reflect a minimal level of engagement with the task.

#### *Level 2*

At level 2, students can generate appropriate ideas for simple visual and written expression tasks as well as those that focus on solving familiar, everyday social problems. With respect to students at level 1, students in level 2 can develop simple written ideas in the form of longer captions or short dialogues. Students at level 2 typically suggest ideas that rely on obvious idea associations for expressive tasks or that refer to existing solutions for problems in social problem-solving tasks. Students can generate more than one appropriate idea for some written expression and social problem-solving tasks, but these ideas are not qualitatively different to one another.

### *Level 3*

At level 3, students can generate one or several appropriate ideas for simple to moderately complex expressive and problem-solving tasks, including extended written ideas that require them to engage and express their imagination and coherently build upon others' ideas. Students at level 3 still typically suggest ideas that rely on obvious idea associations or common themes with respect to their peers, but they begin to demonstrate the ability to recognise and generate original solutions for familiar, everyday problems with a social focus. They may suggest solution ideas that not many other students think of or add an innovative or different twist to more conventional solution ideas.

### *Level 4*

At level 4, students can productively engage in idea generation across a range of expressive and problem-solving tasks. Students at level 4 can also generate original and diverse ideas for simple tasks in more familiar domain contexts. With respect to students at level 3, students at this level can generate an appropriate idea for most types of idea generation task, including more complex or unfamiliar problem-solving tasks and tasks in a scientific context. They can also build on others' ideas for solutions in social and scientific contexts, although they tend to provide an obvious or common iteration with respect to their peers. Students at level 4 can generate their own original ideas in written expression tasks and sometimes when iterating on others' ideas. They can express their imagination in unexpected ways, making unconventional idea associations between elements of the stimulus and their written artefact, or they can add atypical details to elaborate creatively on more common ideas. Students at this level can often suggest two or three qualitatively different ideas in open written expression and social problem contexts but are less successful in more complex or constrained social and scientific problem contexts.

### *Level 5*

At level 5, students can productively engage in creative idea generation, generating both original and diverse ideas for a range of expressive and problem-solving tasks. Students at level 5 can think of qualitatively different ways to express their imagination and to address familiar social and scientific problems. They can make several different idea associations, considering different interpretations and perspectives on the same issue or stimulus. For both simple and more abstract written expression tasks, they can use their imagination to create original written artefacts that make unconventional associations between ideas or that add atypical details to elaborate creatively on common themes. With respect to students at level 4, students can create original visual artefacts that combine elements in an unusual or unexpected way for open visual design tasks. Students at this level can also generate unconventional solution ideas that integrate innovative approaches in familiar social, and sometimes scientific, problem contexts. This includes when tasked to iterate on and improve an existing solution idea in more open, familiar problem contexts.

### *Level 6*

At level 6, students can productively engage in creative idea generation, generating both original and diverse ideas for a wide range of expressive and problem-solving tasks including those in more complex, abstract and unfamiliar contexts. With respect to students at level 5, students at this level can identify weaknesses in existing solutions to social or scientific problems, including those that are in less familiar contexts, and build on this understanding to suggest original and innovative ways to improve solutions. They can also generate several appropriate solution ideas for complex social and scientific problems that require more specific knowledge of the domain context and that have a more restricted solution space. For expressive tasks, students at level 6 can create and improve more abstract visual designs, combining

visual elements and representations in unexpected ways and conveying an original interpretation or iteration of an existing representation.

### ***Cutpoints defining the proficiency levels for creative thinking***

Annex Table 18.A.3 presents the cut points used to assign items and students to a proficiency level for the creative thinking assessment. As with the other PISA domains (see Chapter 17), values in the table are the lower bound for the corresponding level. For example, Level 6 begins with 48.00. Level 5 begins with 41.00 and ends just below 48.00 (i.e. 47.99), where Level 6 begins. Below Level 1 are those with values lower than 6.00. In other words, those reaching a level are those with a score or difficulty at or above the given cut point. This same interpretation applies to all proficiency scales used in PISA. Annex Table 18.A.4 presents a mapping of the released items to the different levels of the proficiency scales.

## References

- ACT (2020), *PISA 2022 Creative Thinking Limited Field Trial Research Report*, ACT, Iowa City, IA. [8]
- ACT (2019), *PISA 2021 Creative Thinking Validation Study Research Report*, ACT, Iowa City, IA. [7]
- ACT (2018), *PISA 2021 Creative Thinking Cognitive Lab Research Report*, ACT, Iowa City, IA. [6]
- Cropley, A. (2006), "In Praise of Convergent Thinking", *Creativity Research Journal*, Vol. 18/3, pp. 391-404. [5]
- Guilford, J. (1956), "The structure of intellect", *Psychological Bulletin*, Vol. 53/4, pp. 267-293, <https://doi.org/10.1037/h0040755>. [2]
- Guilford, J. (1950), "Creativity", *American Psychologist*, Vol. 5/9, pp. 444-454, <https://doi.org/10.1037/h0063487>. [4]
- OECD (2023), *PISA 2022 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/dfe0bf9c-en>. [1]
- Plucker, J., R. Beghetto and G. Dow (2004), "Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research", *Educational Psychologist*, Vol. 39/2, pp. 83-96. [3]

# Annex 18.A. Development and Validation of the Creative Thinking Assessment in PISA 2022

**Annex Table 18.A.1. Overview of Creative Thinking Test Metrics and Distribution in PISA 2022**

Tables	Title
Table 18.A.2	Distribution of items across facets and domains for the PISA 2022 creative thinking test
Table 18.A.3	Cutpoints for the Creative Thinking Scale
Table 18.A.4	A map for released creative thinking items

**Annex Table 18.A.2. Distribution of items across facets and domains for the PISA 2022 creative thinking test**

Domain	Facet					
	Field trial			Main survey		
	Generate diverse ideas	Generate creative ideas	Evaluate and improve ideas	Generate diverse ideas	Generate creative ideas	Evaluate and improve ideas
Written expression	4	6	2	4	6	2
Visual expression	2	2	4	1	1	2
Social problem solving	4	3	3	4	3	3
Scientific problem solving	4	1	3	3	1	2
Total	14	12	12	12	11	9

**Annex Table 18.A.3. Cutpoints for the Creative Thinking Scale**

Cutpoint	Level name
48.00	Level 6
41.00	Level 5
32.00	Level 4
23.00	Level 3
15.00	Level 2
6.00	Level 1

**Annex Table 18.A.4. A map for released creative thinking items**

Level	Cutpoint	Item	Item difficulty
Level 6	48.00	Science Fair Poster (DT200Q02C2) – Full credit	56.66
		Science Fair Poster (DT200Q01C2) – Full credit	53.91
		Library Accessibility (DT500Q02C2) – Full credit	53.35
		Save the River (DT690Q02C2) – Full credit	49.59
Level 5	41.00	<i>Library Accessibility (DT500Q02C2) – Partial credit</i>	46.73
		Save the River (DT690Q01C)	46.41
		Carpooling (DT630Q01C2) – Full credit	45.14
		Illustration Titles (DT300Q01C2) – Full credit	44.65



		Save the Bees (DT400Q02C2) – Full credit	43.69
		Space Comic (DT240Q01C2) – Full credit	42.92
Level 4	32.00	<i>Carpooling (DT630Q01C2) – Partial credit</i>	39.40
		2983 (DT370Q01C2) – Full credit	37.56
		Library Accessibility (DT500Q01C) – Full credit	37.00
		<i>Save the River (DT690Q02C2) – Partial credit</i>	36.63
		<i>Save the Bees (DT400Q02C2) – Partial credit</i>	36.14
Level 3	23.00	Robot Story (DT570Q01)	31.09
		2983 (DT370Q01C2) – Partial credit	27.18
Level 2	15.00	<i>Library Accessibility (DT500Q01) – Partial credit</i>	19.02
		<i>Space Comic (DT240Q01C2) – Partial credit</i>	18.50
Level 1	6.00	<i>Science Fair Poster (DT200Q02C2) – Partial credit</i>	14.59
		<i>Science Fair Poster (DT200Q01C2) – Partial credit</i>	11.80
		<i>Illustration Titles (DT300Q01C2) – Partial credit</i>	6.74

# 19

## Scaling procedures and construct validation of context questionnaire data

### Introduction

The PISA 2022 Context Questionnaires are based on the questionnaire framework (OECD, 2023<sup>[1]</sup>) described in Chapter 5 of this technical report. Many questionnaire items were designed to be combined in some way in order to represent latent constructs that cannot be observed directly (e.g., a student's mathematics self-efficacy; sense of belonging; or economic, social, and cultural status). To construct meaningful indices, transformations or scaling procedures were applied to these items.

In the following sections, these indices are referred to as *derived variables* (DVs). This chapter describes the DVs based on one or more items that were constructed and validated for all questionnaires across the respondent groups – students, parents, schools, and teachers – administered in PISA 2022.

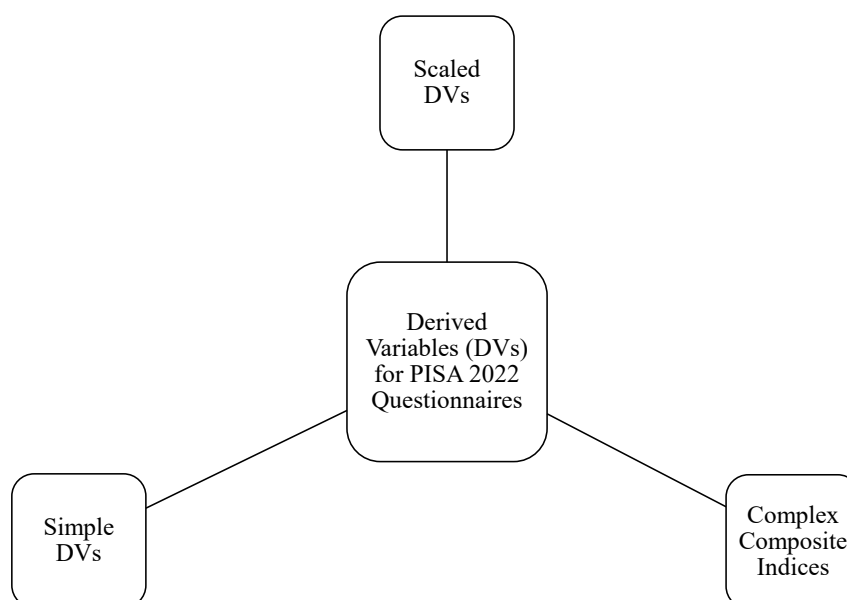
As in the previous PISA surveys, three different kinds of DVs can be distinguished (see Figure 19.1):

- *simple questionnaire indices* constructed through the arithmetical transformation or recoding of one or more items;
- *scaled indices* based on item response theory (IRT) scaling; and
- *complex composite indices* based on a combination of two or more indices.

As described in Chapter 5, the PISA 2022 Context Questionnaires included a broad scope of contextual factors assessed with different questionnaire instruments. While the student and school questionnaires were mandatory in all countries/economies, many countries/economies also administered an optional questionnaire for the parents of the participating students. In addition, countries/economies could choose to administer the optional Financial Literacy Questionnaire, the Information and Communication Technology (ICT) Familiarity Questionnaire, and the Well-Being Questionnaire to students. Moreover, several countries/economies also chose to administer the optional Teacher Questionnaire, which included questionnaires for mathematics teachers and general teachers.

This chapter describes the methodology used for the scaled DVs and also presents an overview of all the simple and scaled DVs for each questionnaire.

Figure 19.1. Types of derived variables for questionnaires in PISA 2022



### ***Within-construct matrix sampling***

Previous PISA cycles have used different strategies for collecting data on relevant contextual variables via the Student Questionnaire. For example, PISA 2012 used a three-form booklet design through which each student was administered items for some but not all of the constructs in the questionnaire. The main benefit of this design was that it allowed for the collection of data on approximately 33% more contextual items at the population level without overburdening students with a single-booklet design. However, a disadvantage of this design was the introduction of systematic missing data for students at the construct level, preventing researchers conducting secondary analyses to fully study the relationships between all possible sets of constructs, since no student was administered items for all of the constructs. PISA 2015 and PISA 2018 used a single-form booklet design through which all students were administered the same items. The main benefit of this design was that it allowed for the creation of a database without systematic missing data at the construct or item level, enabling researchers conducting secondary analyses to study the relationships between all possible sets of items and constructs. However, a disadvantage of this design was that it only allowed for the administration of a smaller set of items for each construct compared to the design used in PISA 2012, leading to a large number of relatively short 3-item scales with somewhat limited representation of the broad underlying construct.

PISA 2022 used a new within-construct matrix sampling design that combined the advantages of the multi-form and single-form booklet designs. This design was studied extensively using data from previous PISA cycles as well as the Field Trial data before it was implemented in the Main Survey (Bertling and Weeks, 2018<sup>[2]</sup>; 2020<sup>[3]</sup>; Bertling et al., 2020<sup>[4]</sup>). Specifically, with this new design, every student was administered a random subset of five items for each construct. This design ensured that each item was administered to approximately the same number of students in each country/economy as well as the overall sample. It also allowed each construct to be assessed in larger breadth, kept individual students' burden comparable to previous cycles, and substantially reduced the reading load for students by displaying only five items on each screen.

This within-construct matrix design was only used for the IRT based scales in the Student Questionnaire, as the primary reporting objective for these scales was at the construct level instead of the item level. Also,

this design was not used for any scales pertaining to the economic, social, and cultural status index. In addition, it was not used for any of the optional questionnaires administered to students (i.e., Financial Literacy Questionnaire, ICT Familiarity Questionnaire, Well-Being Questionnaire) or questionnaires administered to adult respondents (i.e., Parent Questionnaire, School Questionnaire, Teacher Questionnaire) due to the smaller sample sizes for these questionnaires. Table 19.1 provides a list of the 32 scales in the Student Questionnaire that were administered using the within-construct matrix sampling design.

## Scaling methodology and reporting of scores

### **Scaling methodology**

As in previous cycles of PISA, some of the DVs were constructed using IRT. More specifically, the two-parameter logistic model (2PLM) (Birnbaum, 1968<sup>[5]</sup>) was used to scale items with only two response categories (i.e., dichotomous items), while the generalised partial credit model (GPCM) (Muraki, 1992<sup>[6]</sup>), was used to scale items with more than two response categories (i.e., polytomous items).<sup>1</sup> A detailed explanation of each model is in the following sections. The software mdltm (version 1.965) (Shin et al., 2017<sup>[7]</sup>; von Davier, 2015<sup>[8]</sup>) was used for the scaling.

In the initial scaling, item parameters were estimated using data from all individuals with available data from all participating countries/economies. Each country/economy was included in the analysis using a senate weight (SENWT). The senate weight is a linear transformation of the student full sampling weight (W\_FSTUWT) such that the sum of SENWT for all cases within a country/economy add up to a constant of 5 000. Due to missing responses within each country/economy, the sum of the SENWT of the cases used in the calibration of each scale varied on a scale-by-scale basis.

For countries/economies with more than one language group, a language group was treated as an independent group in the scaling process if the group's sample size was over 150 and the sum of the weights was over 300. The groups used in the scaling are called *country-by-language groups* since they are defined by both country/economy and language group. For simplicity, the country-by-language groups are also called *groups* in the remainder of this chapter. Note that if the sample size for an entire country/economy was 150 or less, data from the country/economy were not included in the estimation of the item parameters, and the country/economy was assigned international item parameters (explained below) that had been estimated with data from the other countries/economies.

Several of the scales had items with negative valence. These are items for which a higher response category signified a lower level of the construct being measured, and vice versa. The responses to these items were reverse-coded prior to scaling. For all items, including the reverse-coded items, the responses were recoded so that the response corresponding to the lowest level of the construct was coded as 0. Any missing response data, whether it was because an item was not administered or a student did not respond to an item, were ignored and were not included in the analysis.

### **The two parameter logistic model (2PLM)**

The 2PLM, a generalisation of the Rasch model (Rasch, 1960<sup>[9]</sup>), assumes that the probability of a response  $x$  to be positive (coded as 1 in this case) by individual  $v$  to item  $i$  depends on the difference between the respondent  $v$ 's trait level  $\theta$  and the location of the item  $\beta$ . In addition, the 2PLM postulates that for every item, the association between this difference and the response probability depends on an additional item discrimination parameter  $\alpha$ . The equation for the response probability for an item under the 2PLM is presented in Formula 19.1:

**Formula 19.1**

$$P(x_{iv} = 1 | \theta_v, \beta_i, \alpha_i) = \frac{\exp(D\alpha_i(\theta_v - \beta_i))}{1 + \exp(D\alpha_i(\theta_v - \beta_i))}$$

The item location parameter  $\beta$  can be regarded as the item's general location on the latent continuum of the construct being measured. Items with a higher  $\beta$  parameter require a higher latent trait for a positive response to be selected.

The item discrimination parameter  $\alpha$ , which was scaled by a constant  $D = 1.7$  starting in PISA 2015 when the 2PLM was used instead of the Rasch model, characterises how quickly the probability of responding positively to an item approaches 1 with an increase in the trait level  $\theta$ . In other words,  $\alpha$  describes how well a certain item relates to the latent trait  $\theta$  and, therefore, discriminates between individuals with different trait levels. To solve the indeterminacy of the IRT scale, the average of the item discrimination parameters  $\alpha$  across all the items in the scale was constrained to 1. A special case of the 2PLM is when  $\alpha = 1$  for all items, in which case the model is equivalent to the Rasch model.

**The generalised partial credit model (GPCM)**

The GPCM (Muraki, 1992<sup>[6]</sup>) is a mathematical model for the probability that an individual will select a certain response category for an item with more than two response categories. Note that the GPCM is a generalisation of the 2PLM and that it reduces to the 2PLM when applied to items with only two response categories. For an item  $i$  with  $m + 1$  ordered categories, the probability of an individual selecting a certain response category  $k$  ( $0, 1, 2, \dots, m$ ) under the GPCM and adopting the same notation employed above can be written as:

**Formula 19.2**

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

As with the 2PLM, the overall item location parameter  $\beta$  can be regarded as the item's general location on the latent continuum of the construct being measured. Items with a higher  $\beta$  parameter require a higher latent trait for a higher response category to be selected.  $d$  is the step parameter (of which there are  $m$  for an item with  $m + 1$  categories, with the step parameters for each item summing to 0) which represents the deviation of the category intersection  $\delta$  from the general location  $\beta$ .

The category intersection  $\delta$  is the intersection between two neighbouring category characteristic curves, in other words, the point on the latent continuum  $\theta$  at which a higher response category is more likely to be selected (e.g., when the individual is more likely to select "disagree" than "strongly disagree"). Note that  $\beta$  and  $d$  can be used to calculate the category intersection  $\delta$  using Formula 19.3.

**Formula 19.3**

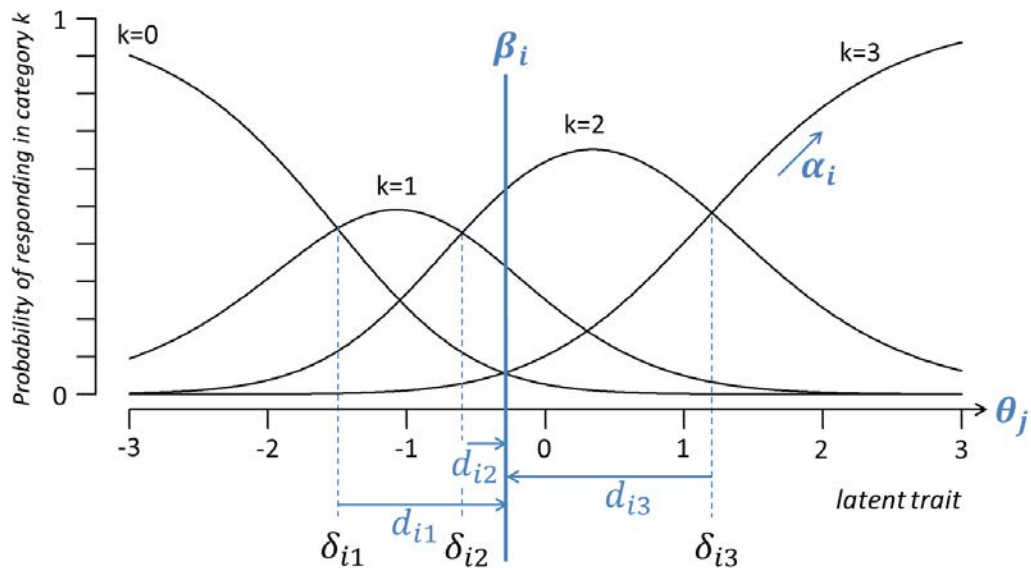
$$\delta_k = \beta - d_k$$

The discrimination parameter  $\alpha$ , which was scaled by a constant  $D = 1.7$  starting in PISA 2015, signifies the slope of the category characteristic curves. In other words, it indicates how well selecting a certain response category discriminates between individuals on the latent continuum  $\theta$ . To solve the indeterminacy of the IRT scale, the average of the item discrimination parameters  $\alpha$  across all the items in the scale was

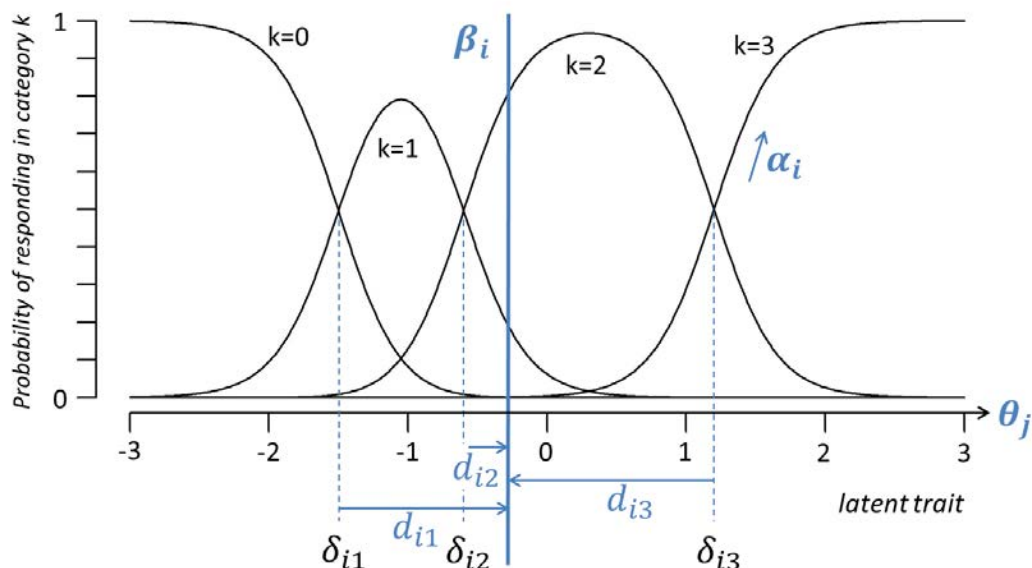
constrained to 1. A special case of the GPCM is when  $\alpha = 1$  for all items, in which case the model is equivalent to the partial credit model (PCM) (Masters, 1982<sup>[10]</sup>).

Figure 19.2 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”), with the three item parameters used in the GPCM (i.e.,  $\alpha$ ,  $\beta$  and  $d$ ) represented in the figure. For comparison, Figure 19.3 displays the category characteristic curves of an item for which only the  $\alpha$  parameter has been increased while the  $\beta$  and  $d$  parameters were kept the same as in Figure 19.2.

**Figure 19.2. Category characteristic curves for a four-category item under the generalised partial credit model (GPCM)**



**Figure 19.3. Illustration of how an increase in the slope parameter  $\alpha$  affects the category characteristic curves of the model above**



### **Special handling of trend scales**

For the trend scales, the scaling process began by fixing the item parameters of the trend items to the parameters that had been estimated for each group in the previous cycle, a procedure called fixed parameter linking. Also, in line with the models that were used in the past cycles, the trend scales linked to PISA 2018 were scaled using the 2PLM and GPCM, while the trend scales linked to PISA 2012 were scaled using the Rasch model and PCM. This was done so that the scale scores from the current cycle would be comparable to the scale scores from the previous cycle. To compute trends, a scale needed to have at least three trend items, but some trend scales consisted of both trend items and new items. In this case, the item parameters for the trend items were fixed at the beginning of the scaling process, but the item parameters for the new items were estimated using the PISA 2022 data. Note that all the items in the trend scales were also evaluated for the goodness-of-fit of the trend parameters, a process described below. Please see Table 19.2 for a full list of trend scales in PISA 2022.

### **Releasing item parameters**

PISA 2022 adopted and further refined the approach for evaluating the invariance of latent constructs across groups using multiple-group concurrent calibration with partial invariance constraints, a method which was first introduced in PISA 2015.

As explained above, in the initial scaling, item parameters were estimated using data from all individuals with available data from all countries/economies. The item parameters that were estimated in this initial scaling process are called *international parameters* since they were estimated using responses from most or all participating countries/economies. After the initial scaling, the fit of the international parameters for each item was evaluated for each group using the root mean square deviation (RMSD). The  $RMSD_g$  for group  $g$  is defined as:

## Formula 19.4

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

quantifying the difference between the observed item characteristic curve (ICC) for the group based on the pseudo counts from the E-step of the Expectation-Maximisation (EM) algorithm ( $P_{obs,gk}(\theta)$ ) with the model-based Item Characteristic Curve ICC ( $P_{exp,gk}(\theta)$ ) (Shin et al., 2017<sup>[7]</sup>). RMSD values range from 0 to 1, with values close to 0 indicating good item fit, meaning that the model-based item parameters fit the data for the group well. Note that the RMSD statistic is sensitive to group-specific deviations of both the item location parameter  $\beta$  and the item discrimination parameter  $\alpha$ .

When the RMSD for an item\*group exceeded a pre-defined cut-off of 0.25, it was considered that the model-based item parameters did not fit the group's data well, and unique item parameters (also called *group-specific item parameters*) were estimated for the group using data only from that group. However, if more than one group had similar response patterns for an item, data from those groups were pooled together and the same unique parameters were estimated for those groups. This process is called *releasing item parameters*. Item parameters were released until all item\*groups had an RMSD value under 0.25.

In PISA 2015 and PISA 2018, an RMSD value of 0.3 was used as the cut-off criterion for releasing item parameters for the context questionnaires. However, an analysis of the scaling results from the PISA 2018 Student Questionnaire suggested that this threshold may have over-emphasised international comparability of the model over group-level model-data fit, as very few item\*groups received unique parameters. For PISA 2022, the RMSD threshold for releasing item parameters for the context questionnaires was lowered to 0.25, as it was found that this new threshold could improve the group-level model-data fit without weakening the comparability of the model across groups (as measured by the percent of item\*groups with unique parameters, number of groups with international parameters for three or more items in a scale, and the rank order correlation of the scale scores from the models with and without unique parameters).

The final distribution of the RMSD values across groups for each scale item after the final scaling is documented in Annex E. Note that the figures do not include RMSD values for item\*groups with an unweighted sample size of 150 or less, as the sample size was too small to estimate stable group-specific ICCs.

### Scale scores

The scaling process described above produced weighted likelihood estimates (WLE) (Warm, 1989<sup>[11]</sup>) for each individual. These WLE scores were subsequently standardised through the process described in the following sections. Note that if an individual had fewer than three valid responses for a scale, a WLE score was not produced for the individual and his/her scale score was replaced with “99” in the SPSS file and “.M” in the SAS file.

#### *New scales*

For the new scales, the original WLE scores were transformed into a reporting metric to have a mean of 0 and a standard deviation of 1 across the OECD countries, using senate weights for all cases with available data. The transformation was achieved by applying Formula 19.5:



## Formula 19.5

$$\theta'_v = \frac{\theta_v - \bar{\theta}_{OECD}}{\sigma_{\theta(OECD)}}$$

where  $\theta'_v$  is the scale score on the reporting metric,  $\theta_v$  is the original WLE,  $\bar{\theta}_{OECD}$  is the mean of the original WLEs across the OECD countries, and  $\sigma_{\theta(OECD)}$  is the standard deviation of the original WLEs across the OECD countries. The transformation constants that were used to transform the original WLEs into the reporting scale are displayed in Table 19.3.

For the new scales, an average scale score of 0 is expected when calculated across all OECD countries using the senate weights. A negative scale score does not imply that a student responded negatively to the items in the scale. Rather, it means that the student is below the OECD average.

### *Trend scales*

For the trend scales, to ensure the comparability of the scale scores from the current cycle to the scale scores from the previous cycle, the original WLEs of PISA 2022 were transformed using the same transformation constants of the original WLEs from the cycle to which the current cycle was linked. Table 19.4 presents the transformation constants of the original WLEs in PISA 2018 for the trend scales linked to PISA 2018, while Table 19.5 presents the transformation constants used in PISA 2012 for the trend scales linked to PISA 2012.

### **Criteria for suppressing scale scores**

The scale scores of individuals or groups were suppressed under the following conditions.

#### *Low internal consistency*

Cronbach's alpha coefficient was used to check the internal consistency of each scale for each group. This coefficient ranges from 0 to 1, with a higher value indicating higher internal consistency. A group needed to have a Cronbach's alpha of at least 0.60 for a scale in order for the group's scale scores to be reported. Scale scores were suppressed for countries/economies in which one or more language groups had a Cronbach's alpha under 0.60 for the scale.

#### *Few items with international parameters*

For each scale, a group needed to have at least three items with international parameters in order for the scale scores of the group to be considered comparable to the scale scores of the other groups. Scale scores were suppressed for countries/economies in which one or more language groups had less than three items with international parameters for the scale. The scale scores for the individuals in these countries/economies were replaced with "97" in the SPSS file and ".N" in the SAS file.

#### *Lack of trend items with international parameters*

For the trend scales, a group needed to have at least three trend items with international parameters in order for the PISA 2022 scales scores for the group to be considered comparable to the scale scores of the previous cycle to which the current cycle was linked. Scale scores were suppressed for countries/economies in which one or more language groups had less than three trend items with international parameters for the scale. The scale scores for the individuals in these countries/economies were replaced with "97" in the SPSS file and ".N" in the SAS file.

## Student Questionnaire derived variables

There were 86 variables derived from the Student Questionnaire, including 43 simple DVs, 42 IRT scaled DVs, and one complex composite index. The DVs are shown in Table 19.6 and will be described in the following sections; the first section covers all simple DVs, the second section covers those that are based on IRT scaling, and the last section covers the complex composite index. The simple and scaled DVs are organised first by framework module (please see Chapter 5 for a description of modules) and then alphabetical order within modules.

### Simple questionnaire indices

#### *Basic demographics (Module 1)*

##### **Student's age (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, which was obtained from school records from the student sampling data and validated by comparing to the students' responses in the questionnaire. Data on students' age were obtained from both the questionnaire (ST003) and the student tracking forms. The formula for computing AGE was:

#### Formula 19.6

$$AGE = (100 + T_y - S_y) + (T_m - S_m)/12$$

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.

##### **Grade compared to modal grade in country (GRADE)**

The relative grade index (GRADE) was computed to capture between-country/economy variation. It indicates whether students are in the country/economy's modal grade (value of 0), or the number of grades below or above the modal grade in the country. The information about the students' grade level was obtained from school records from the student sampling data and validated by comparing the students' responses in the Student Questionnaire (ST001).

##### **Gender (ST004D01T)**

The gender of a student which was obtained from school records from the student sampling data and validated by comparing to the student's responses in the questionnaire (ST004).

#### *Economic, social and cultural status (Module 2)*

##### **Mother's level of education (MISCED)**

Student responses to questions ST005 and ST006 regarding their mothers' education were used to derive the mother's level of education (MISCED) index, where education level ranged from "1" less than ISCED level 1 to "10" ISCED level 8, as noted in Table 19.7.<sup>2</sup>

### **Father's level of education (FISCED)**

Student responses to questions ST007 and ST008 regarding their fathers' education were used to derive the father's level of education (FISCED) index, where education level ranged from "1" less than ISCED level 1 to "10" ISCED level 8, as noted in Table 19.7 above.

### **Highest level of education of parents (HISCED)**

Students' responses to questions ST005, ST006, ST007, and ST008 regarding their mothers' and fathers' education were used to derive the index of highest education level of parents (HISCED). The index is equal to the highest ISCED level of either parent.

### **Highest education of parents in years (PAREDINT)**

The index of the highest education of parents in years, PAREDINT, was based on the median cumulative years of education associated with completion of the highest level of parental education (HISCED). Cumulative years of education values used in PISA 2018 were assigned to each ISCED level (see Table 19.7). Mother's occupational code (OCOD1)

Students' responses to the fill-in question ST014 about their mothers' occupation were human-coded based on the International Standard Classification of Occupations (ISCO)-08 classification system, resulting in the mother's occupational code (4-digit ISCO; ILO, 2007) index, OCOD1. These 4-digit codes range from 0000 to 9705. Codes 0000 to 9629 are occupations from the ISCO-08 classification system. Codes 9701-9705 were used to classify responses that fell outside of the ISCO-08 classification system. Specifically, the code 9701 indicates "stay-at-home parent", 9702 indicates "student", and 9703 indicates "social beneficiary (e.g., unemployed, retired, sick)". Lastly, "I don't know" responses were coded 9704 and vague responses (e.g., a good job, a well-paid job) were coded 9705.

### **Father's occupational code (OCOD2)**

Students' responses to the fill-in question ST015 about their fathers' occupation were human-coded based on the ISCO-08 classification system, resulting in the father's occupational code (4-digit ISCO) index, OCOD2. These 4-digit ISCO-08 codes range from 0000 to 9705. Codes 0000 to 9629 are occupations from the ISCO-08 classification. Codes 9701-9705 were used to classify responses that fell outside of the ISCO-08 classification. Specifically, the code 9701 indicates "stay-at-home parent", 9702 indicates "student", and 9703 indicates "social beneficiary (e.g., unemployed, retired, sick)". Lastly, "I don't know" responses were coded 9704 and vague responses (e.g., a good job, a well-paid job) were coded 9705.

### **Mother's occupational status (BMMJ1)**

The mother's occupational status index, BMMJ1, was derived from the OCOD1 index and international socio-economic index of occupational status (ISEI) (Ganzeboom and Treiman, 2003<sub>[12]</sub>) scores. The 4-digit ISCO-08 occupation codes in OCOD1 were mapped onto ISEI ratings.

### **Father's occupational status (BFMJ2)**

The father's occupational status index, BFMJ2, was derived from the OCOD2 index and international socio-economic index of occupational status (ISEI) scores. The 4-digit ISCO-08 occupation codes in OCOD2 were mapped onto ISEI occupational status scores.

### **Highest parental occupational status (HISEI)**

This highest parental occupational status index (HISEI) was based on the 4-digit ISCO-08 occupational codes that were human coded from students' responses to questions ST014 and ST015 about their mother and father's occupations, respectively. The index was equal to the higher of the mother's (BMMJ1) and father's (BFMJ2) ISEI scores.

#### *Educational pathways and post-secondary aspirations (Module 3)*

### **Duration in early childhood education and care (DURECEC)**

Questions ST125 and ST126 measure the starting age in ISCED 1 and ISCED 0. The indicator DURECEC is built as the difference of ST126 and ST125 plus the value of "2" to indicate the number of years a student spent in early childhood education and care.

### **Study programme level and orientation (ISCEDP)**

PISA collects data on study programmes available to 15-year-old students in each country/economy. This information is obtained through the student tracking form and the Student Questionnaire (ST002). In the final database, all national programmes are included in a separate DV (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 2011).

The study programme level and orientation index (ISCEDP) is a three-digit index that describes whether students were at the lower or upper secondary level and (ISCED 2 or ISCED 3) and whether their programmes were general or vocational and sufficient for level completion with direct access to tertiary or post-secondary non-tertiary education. ISCEDP values and labels can be found in Table 19.8.

### **Grade repetition (REPEAT)**

Students' answers on question ST127 of whether and, if yes, how often they have ever repeated a grade at ISCED levels 1, 2, and 3 were combined into the index REPEAT. Each item included three response options ("No, never", "Yes, once", "Yes, twice or more"). REPEAT took the value of "0" if the student never repeated a grade (student did not select options 2 or 3 for any of the three items) and the value of "1" if the student repeated a grade at least once (student selected options 2 or 3 for at least one of the three items). The index was assigned a missing value if none of the three response options were selected in any levels.

### **Missing school (MISSSC)**

Students' answers on question ST260 of whether and, if yes, how often they have ever missed school for more than three months in a row at ISCED levels 1, 2, and 3 were combined into the index MISSSC. Each item included three response options ("No, never", "Yes, once", "Yes, twice or more"). MISSSC took the value of "1" if the student selected options 2 or 3 for at least one of the three items, and the value "0" otherwise. The index was assigned a missing value if none of the three response options were selected in any levels.

### **Skipping classes or days of school (SKIPPING)**

Students' responses to whether, in the two weeks prior to the PISA test, they had skipped classes (ST062Q02TA) or days of school (ST062Q01TA) at least once were used to derive an indicator of student truancy. Both questions have four response options ("Never", "One or two times", "Three or four times", "Five or more times"). The indicator takes a value of 0 if students reported that they had not skipped any

class or day of school in the two weeks before the PISA test, and a value of 1 if students reported that they had skipped classes or days of school at least once in the same period.

### **Arriving late for school (TARDYSD)**

Students responded to a question about whether and how frequently they had arrived late for school during the two weeks prior to the PISA test (ST062Q03TA). TARDYSD takes a value of “0” for on-time students if students reported that they had not arrived late for school, a value of “1” for occasional late arrivals if students report they arrived late for school one or two times, and “2” for frequent late arrivals if students reported they had arrived late for school three or more times.

### **Highest expected educational level (EXPECEDU)**

Students’ responses which of a list of possible educational levels they expect to complete in question ST327 were transformed into the index of “Highest Expected Educational Level”. This DV has been newly created for 2022. Values on the index can range from “Less than ISCED level 2” to “ISCED level 8”. Scores are assigned as shown in Table 19.9.

### **Expected occupation (OCOD3) and Expected occupation status (BSMJ)**

Students’ responses to the fill-in question ST329 about what kind of job they expect to have when they are about 30 years old were human-coded based the ISCO-08 classification system, resulting in the index “Expected Occupation (OCOD3)”. These ISCO codes were then mapped to the international socio-economic index of occupational status (ISEI) (Ganzeboom and Treiman, 2003<sup>[12]</sup>) in variable BSMJ. Higher scores on this variable indicate higher levels of a student’s expected occupational status.

### **Clear idea about future job (SISCO)**

The students who had a clear idea about their future job index (SISCO) was based on the human-coded open-ended expected occupation index, OCOD3, which was derived from question ST329. Students who had no clear idea about their future jobs were considered those who indicated “I do not know” or gave a vague answer such as “a good job”, “a quiet job”, “a well-paid job”, “an office job” in response to question ST329. In the OCOD3 index, “I don’t know” responses were coded 9704 and vague responses were coded 9705. Examples of invalid responses include students who did not answer the question or gave an answer, such as a smiley face. Specifically, a value of “0” is assigned on the index if OCOD3 values are 9704 or 9705, and a value of “1” is assigned if OCOD3 values are 0000 to 9703.

### *Migration and language exposure (Module 4)*

Based on students’ responses to question ST019 (“In what country were you and your parents born?”), five indices are created as outlined below.

#### **Student’s country of birth (COBN\_S)**

This index has the value “1” if the student selected the country of test (“Country A”) in question ST019AQ01T, and “0” otherwise.

#### **Student mother’s country of birth (COBN\_M)**

This index has the value “1” if the student selected the country of test (“Country A”) in question ST019BQ01T, and “0” otherwise.

### **Student father's country of birth (COBN\_F)**

This index has the value “1” if the student selected the country of test (“Country A”) in question ST019CQ01T, and “0” otherwise.

### **Index on immigrant background (IMMIG)**

The index on immigrant background (IMMIG) is calculated from the three variables above (COBN\_S, COBN\_M, COBN\_F), and has the categories as listed below. Students with missing responses for either the student or for both parents were given missing values for this variable.

1. Native students (those students who had at least one parent born in the country/economy);
2. Second-generation students (those born in the country/economy of assessment but whose parent[s] were born in another country/economy);
3. First-generation students (those students born outside the country/economy of assessment and whose parents were also born in another country/economy).

### **Language spoken at home (LANGN)**

Students also indicated what language they usually spoke at home, and the database includes a variable (LANGN) containing country/economy-specific code for each language.

### *Subject-specific beliefs, attitudes, feelings and behaviours (Module 7)*

#### **Relative motivation to do well in mathematics compared to other core subjects (MATHMOT)**

This simple index captures whether students indicate being more motivated to do well in mathematics than in Test Language and Science class. If students endorsed question ST268Q07JA (“I want to do well in my mathematics class.”) stronger than both items ST268Q08JA (“I want to do well in my <test language> class.”) and ST268Q09JA (“I want to do well in my <science> class.”), they received a “1” on this index, otherwise “0”. Please note that this index captures students’ relative motivation for math rather than their absolute motivation for mathematics. The latter is captured by the original response to the item.

#### **Perception of mathematics as easier than other core subjects (MATHEASE)**

This simple index captures whether students indicate they perceive mathematics as easier compared to the Test Language and Science. If students endorsed question ST268Q014A (“Mathematics is easy for me.”) stronger than both items ST268Q05JA (“<Test language> is easy for me.”) and ST268Q06JA (“<Science> is easy for me”), they received a “1” on this index, otherwise “0”. Please note that this index captures students’ relative easiness of math rather than their absolute easiness rating for mathematics. The latter is captured by the original response to the item.

#### **Preference of mathematics over other core subjects (MATHPREF)**

This simple index captures whether students indicate they preferred mathematics over Test Language and Science. If students endorsed question ST268Q01JA (“Mathematics is one of my favourite subjects”) stronger than both items ST268Q02JA (“<Test language> is one of my favourite subjects.”) and ST268Q03JA (“<Science> is one of my favourite subjects.”), they received a “1” on this index, otherwise “0”. Please note that this index captures students’ relative preference of math rather than their absolute preference for mathematics. The latter is captured by the original response to the item.

### *Out-of-school experiences (Module 10)*

#### **Exercising or practising a sport before or after school (EXERPRAC)**

Students' answers on how many days during a typical school week they exercised or practised a sport before going to school and/or after leaving school in questions ST294 and ST295 were scaled into the index of "Exercise or practise a sport before or after school". Each item included six response options ("0 days", "1 day", "2 days", "3 days", "4 days", "5 or more days"). Values on this index range from 0 (no exercise or sports) to 10 (10 or more times exercise or sport a per week).

#### **Studying for school or homework before or after school (STUDYHMW)**

Students' answers on how many days during a typical school week they studied for school or homework before going to school and/or after leaving school in questions ST294 and ST295 were scaled into the index of "Study for school or homework before or after school". Each item included six response options ("0 days", "1 day", "2 days", "3 days", "4 days", "5 or more days"). Values on this index range from 0 (no studying) to 10 (10 or more times of studying per week).

#### **Working for pay before or after school (WORKPAY)**

Students' answers on how many days during a typical school week they worked for pay before going to school and/or after leaving school in questions ST294 and ST295 were scaled into the index of "Work for pay before or after school". Each item included six response options ("0 days", "1 day", "2 days", "3 days", "4 days", "5 or more days"). Values on this index range from 0 (no work for pay) to 10 (10 or more times of working for pay per week).

#### **Working in household or taking care of family members (WORKHOME)**

Students' answers on how many days during a typical school week they worked in the household or took care of a family member before going to school and/or after leaving school in questions ST294 and ST295 were scaled into the index of "Work in household or take care of family members". Each item included six response options ("0 days", "1 day", "2 days", "3 days", "4 days", "5 or more days"). Values on this index range from 0 (no work in household or care of family members) to 10 (10 or more times of working in household or caring for family members per week).

### ***Derived variables based on IRT scaling***

The Student Questionnaire provided data for 42 DVs based on IRT scaling. The Cronbach's alpha for each scale and group are presented in Table 19.10, the number of items with international parameters for each scale and group are presented in Table 19.11, the number of trend items with international parameters for each trend scale and group are presented in Table 19.12, the countries/economies for which the scale scores were suppressed for each scale are presented in Table 19.13, and the groups that did not administer each scale are presented in Table 19.14.

### *Economic, social and cultural status (Module 2)*

#### **Home possessions (HOMEPOS)**

In the HOMEPOS scale (which included questions ST250, ST251, ST253, ST254, ST255, and ST256), students indicated whether their household possessed certain items (e.g., "A room of your own", "Educational software or apps") or how many of an item their household possessed (e.g., "Rooms with a <flush toilet>", "Cars, vans, or trucks"). This scale included 31 items, including four country/economy-

specific items (ST250Q06JA, ST250Q07JA, ST251Q08JA, and ST251Q09JA) that were seen as local measures of family wealth within the country/economy's context.<sup>3</sup> In addition, students answered how many books (ST255) and digital devices with screens (ST253) were in their home. Note that all groups received unique item parameters for the country/economy-specific items (i.e., no international parameters were estimated for these items) and that for some items, the response categories were collapsed to align with the response categories used in previous cycles. Table 19.15 shows the item wording and item parameters for the items in this scale, while Table 19.16 shows how the response categories for each item were recoded prior to scaling.

### **ICT resources (ICTRES)**

Students reported on the availability of 11 Information and Communications Technologies (ICT) resources in their home (e.g., "A computer (laptop, desktop, or tablet) that you can use for school work", "Internet access (e.g., Wi-fi) (excluding through smartphones)") in questions ST250 (which had two response categories), ST253 (which had eight response categories), and ST254 (which had four substantive response categories and an additional response category "I don't know." which was recoded as missing prior to scaling). These items were scaled into the index of "ICT resources". Table 19.17 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded and how the response categories were recoded prior to scaling.

### *Educational pathways and post-secondary aspirations (Module 3)*

#### **Information seeking regarding future career (INFOSEEK)**

Students' ratings of whether they had undertaken a range of possible activities to find out about future study or types of work (e.g., "I did an internship.", "I researched the internet for information about careers.") in question ST330 were scaled into the index of "Information seeking regarding future career". Note that this scale used a within-construct matrix sampling design. Each of the 11 items included in this scale had three response options ("Yes, once", "Yes, two or more times", "No"). Table 19.18 shows the item wording and item parameters for the items in this scale.<sup>4</sup> It also shows how the response categories were recoded prior to scaling.

### *School culture and climate (Module 6)*

#### **Being bullied (BULLIED)**

Students' ratings of how often they had a range of experiences at school that are indicative of being bullied during the past 12 months (e.g., "Other students left me out of things on purpose.", "Other students made fun of me.") in question ST038 were scaled into the index of "Being bullied". Note that this scale was linked to the BEINGBULLIED scale in PISA 2018. Each of the nine items included in this scale had four response options ("Never or almost never", "A few times a year", "A few times a month", "Once a week or more"). Table 19.19 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

#### **Feeling safe (FEELSAFE)**

Students' ratings of their agreement with four statements about their perceived safety (e.g., "I feel safe on my way to school.", "I feel safe in my classrooms at school.") in question ST265 were scaled into the index of "Feeling safe". Each of the four items included in this scale had four response options ("Strongly agree", "Agree", "Disagree", "Strongly disagree"). Table 19.20 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.



### **Mathematics teacher support (TEACHSUP)**

Students' frequency ratings of how often a range of situations occurred in their mathematics lessons (e.g., "The teacher shows an interest in every student's learning.", "The teacher gives extra help when students need it.") in question ST270 were scaled into the index of "Mathematics teacher support". Note that this scale was linked to the TEACHSUP scale in PISA 2012 and was scaled using the PCM, in line with the model used in PISA 2012. Each of the four items included in this scale had four response options ("Every lesson", "Most lessons", "Some lessons", "Never or almost never"). Table 19.21 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling and which items are trend items.

### **Quality of student-teacher relationships (RELATST)**

Students' ratings of their agreement with the eight statements (e.g., "The teachers at my school are respectful towards me.", "When my teachers ask how I am doing, they are really interested in my answer.") in question ST267 were scaled into the index of "Quality of student-teacher relationships". Note that this scale used a within-construct matrix sampling design. Each of the eight items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.22 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### **School safety risks (SCHRISK)**

Students' answers of whether a range of events indicative of safety risks at school occurred during the past four weeks (e.g., "Our school was vandalised.", "I witnessed a fight on school property in which someone got hurt.") in question ST266 were scaled into the index of "School safety risks". Each of the five items included in this scale had two response options ("Yes", "No"). Table 19.23 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### **Sense of belonging (BELONG)**

Students' ratings of their agreement with six statements (e.g., "I feel like I belong at school.", "I feel lonely at school.") in question ST034 were scaled into the index of "Sense of belonging". Note that this scale used a within-construct matrix sampling design and that it was linked to the BELONG scale in PISA 2018. Each of the six items included in this scale had four response options ("Strongly agree", "Agree", "Disagree", "Strongly disagree"). Table 19.24 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling and which items are trend items.

### *Subject-specific beliefs, attitudes, feelings, and behaviours (Module 7)*

### **Growth mindset (GROSAGR)**

Students' ratings of their agreement with a range of statements indicative of their mindset (e.g., "Your intelligence is something about you that you cannot change very much.", "Some people are just not good at mathematics, no matter how hard they study.") in question ST263 were scaled into the index of "Growth mindset". Each of the four items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.25 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### **Mathematics anxiety (ANXMAT)**

Students' ratings of their agreement with statements about a range of attitudes towards mathematics (e.g., "I often worry that it will be difficult for me in mathematics classes.", "I feel anxious about failing in mathematics.") in question ST292 were scaled into the index of "Mathematics anxiety". Note that this scale was linked to the ANXMAT scale in PISA 2012 and was scaled using the PCM, in line with the model used in PISA 2012. Also, it used a within-construct matrix sampling design. Each of the six items included in this scale had four response options ("Strongly agree", "Agree", "Disagree", "Strongly disagree"). Table 19.26 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling and which items are trend items.

### **Mathematics self-efficacy: Formal and applied mathematics (MATHEFF)**

Students' ratings of how confident they felt about having to do a range of formal and applied mathematics tasks (e.g., "Calculating how much more expensive a computer would be after adding tax", "Solving an equation like  $2(x+3) = (x+3)(x-3)$ ") in question ST290 were scaled into the index of "Mathematics self-efficacy: Formal and applied mathematics". Note that this scale was linked to the MATHEFF scale in PISA 2012 and was scaled using the PCM, in line with the model used in PISA 2012. Also, it used a within-construct matrix sampling design. Each of the nine items included in this scale had four response options ("Not at all confident", "Not very confident", "Confident", "Very confident"). Note that in PISA 2012, the response options were presented to the students ordered from "Very confident" to "Not at all confident" possibly eliciting different response patterns related to the format of the question, and not necessarily related to the construct. Because of this, caution should be exercised when comparing scale scores across these two cycles. Table 19.27 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### **Mathematics self-efficacy: Mathematical reasoning and 21st century mathematics (MATHEF21)**

Students' ratings of how confident they felt about having to do a range of mathematical reasoning and 21<sup>st</sup> century mathematics tasks (e.g., "Extracting mathematical information from diagrams, graphs, or simulations", "Using the concept of statistical variation to make a decision") in question ST291 were scaled into the index of "Mathematics self-efficacy: Mathematical reasoning and 21<sup>st</sup> century mathematics". Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had four response options ("Not at all confident", "Not very confident", "Confident", "Very confident"). Table 19.28 shows the item wording and item parameters for the items in this scale.

### **Proactive mathematics study behaviour (MATHPERS)**

Students' frequency ratings of how often they engaged in behaviours indicative of effort and persistence in mathematics (e.g., "I actively participated in group discussions during mathematics class.", "I put effort into my assignments for mathematics class.") in question ST293 were scaled into the index of "Proactive mathematics study behaviour". Note that this scale used a within-construct matrix sampling design. Each of the eight items included in this scale had five response options ("Never or almost never", "Less than half of the time", "About half of the time", "More than half of the time", "All or almost all of the time"). Table 19.29 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### **Subjective familiarity with mathematics concepts (FAMCON)**

Students' ratings of how familiar they were with different mathematical concepts representative of different levels of mathematical skill or understanding (e.g., "Divisor", "Exponential function", "3-dimensional

geometry”) in question ST289 were scaled into the index of “Subjective familiarity with mathematics concepts”. Note that this scale was linked to the FAMCON scale in PISA 2012 and was scaled using the PCM, in line with the model used in PISA 2012. Also, it used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Never heard of it”, “Heard of it once or twice”, “Heard of it a few times”, “Heard of it often”, “Know it well, understand the concept”). Table 19.30 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *General social and emotional characteristics (Module 8)*

All of the scales in this module used a within-construct matrix sampling design and included both positively and negatively valenced items. This allowed us to check the consistency of responses since we would expect those agreeing with the items with positive valence to disagree with items with negative valence, and vice versa. To this effect, some students were identified as extreme straightliners. These were students that were administered items with positive and negative valence and responded to all five items selecting the same extreme response category, “Strongly disagree” or “Strongly agree”. Students that were identified as extreme straightliners were removed from the analysis.

#### **Assertiveness (ASSERAGR)**

Students’ ratings of their agreement with statements about a range of behaviours indicative of assertiveness (e.g., “I take initiative when working with my classmates.”, “I find it hard to influence people.”) in question ST305 were scaled into the index of “Assertiveness”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, “Strongly agree”). Table 19.31 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling. Table 19.32 shows the percent of students in each country/economy that did not receive a scale score for ASSERAGR due to extreme straightlining or, for comparison, for not having enough responses (i.e., less than three responses for the scale). In both cases, the scale scores were replaced with “99” in the SPSS file and “.M” in the SAS file.

#### **Cooperation (COOPAGR)**

Students’ ratings of their agreement with statements about a range of behaviours indicative of cooperation (e.g., “I work well with other people.”, “I get annoyed when I have to compromise with others.”) in question ST343 were scaled into the index of “Cooperation”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, “Strongly agree”). Table 19.33 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling. Table 19.34 shows the percent of students in each country/economy that did not receive a scale score for COOPAGR due to extreme straightlining or, for comparison, for not having enough responses (i.e., less than three responses for the scale). In both cases, the scale scores were replaced with “99” in the SPSS file and “.M” in the SAS file.

#### **Curiosity (CURIOAGR)**

Students’ ratings of their agreement with statements about a range of behaviours indicative of curiosity (e.g., “I like to know how things work.”, “I am more curious than most people I know.”) in question ST301 were scaled into the index of “Curiosity”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, “Strongly agree”). Table 19.35 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to

scaling. Table 19.36 shows the percent of students in each country/economy that did not receive a scale score for CURIOAGR due to extreme straightlining or, for comparison, for not having enough responses (i.e., less than three responses for the scale). In both cases, the scale scores were replaced with “99” in the SPSS file and “.M” in the SAS file.

### **Emotional control (EMOCOAGR)**

Students’ ratings of their agreement with statements about a range of behaviours indicative of emotional control (e.g., “I keep my emotions under control.”, “I get mad easily.”) in question ST313 were scaled into the index of “Emotional control”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, “Strongly agree”). Table 19.37 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling. Table 19.38 shows the percent of students in each country/economy that did not receive a scale score for EMOCOAGR due to extreme straightlining or, for comparison, for not having enough responses (i.e., less than three responses for the scale). In both cases, the scale scores were replaced with “99” in the SPSS file and “.M” in the SAS file.

### **Empathy (EMPATAGR)**

Students’ ratings of their agreement with statements about a range of behaviours indicative of empathy (e.g., “I predict the needs of others.”, “It is difficult for me to sense what others think.”) in question ST311 were scaled into the index of “Empathy”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, “Strongly agree”). Table 19.39 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling. Table 19.40 shows the percent of students in each country/economy that did not receive a scale score for EMPATAGR due to extreme straightlining or, for comparison, for not having enough responses (i.e., less than three responses for the scale). In both cases, the scale scores were replaced with “99” in the SPSS file and “.M” in the SAS file.

### **Perseverance (PERSEVAGR)**

Students’ ratings of their agreement with statements about a range of behaviours indicative of perseverance (e.g., “I keep working on a task until it is finished.”, “I give up after making mistakes.”) in question ST307 were scaled into the index of “Perseverance”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, “Strongly agree”). Table 19.41 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling. Table 19.42 shows the percent of students in each country/economy that did not receive a scale score for PERSEVAGR due to extreme straightlining or, for comparison, for not having enough responses (i.e., less than three responses for the scale). In both cases, the scale scores were replaced with “99” in the SPSS file and “.M” in the SAS file.

### **Stress resistance (STRESAGR)**

Students’ ratings of their agreement with statements about a range of behaviours indicative of stress resistance (e.g., “I remain calm under stress.”, “I get nervous easily.”) in question ST345 were scaled into the index of “Stress resistance”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options (“Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, “Strongly agree”). Table 19.43 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling. Table 19.44

shows the percent of students in each country/economy that did not receive a scale score for STRESAGR due to extreme straightlining or, for comparison, for not having enough responses (i.e., less than three responses for the scale). In both cases, the scale scores were replaced with “99” in the SPSS file and “.M” in the SAS file.

### *Exposure to mathematics content (Module 15)*

#### **Exposure to formal and applied mathematics tasks (EXPOFA)**

Students’ frequency ratings of how often they had encountered a range of formal and applied mathematics tasks during their time at school (e.g., “Calculating how much more expensive a computer would be after adding tax.”, “Solving an equation like  $2(x+3) = (x+3)(x-3)$ ”) in question ST275 were scaled into the index “Exposure to formal and applied mathematics tasks”. Note that this scale used a within-construct matrix sampling design. Each of the nine items included in this scale had four response options (“Frequently”, “Sometimes”, “Rarely”, “Never”). Table 19.45 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

#### **Exposure to mathematical reasoning and 21st century mathematics tasks (EXPO21ST)**

Students’ frequency ratings of how often they had encountered a range of different types of mathematics tasks related to mathematical reasoning and 21<sup>st</sup> century mathematics tasks during their time at school (e.g., “Extracting mathematical information from diagrams, graphs, or simulations”, “Using the concept of statistical variation to make a decision”) in question ST276 were scaled into the index “Exposure to mathematical reasoning and 21<sup>st</sup> century mathematics tasks”. Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had four response options (“Frequently”, “Sometimes”, “Rarely”, “Never”). Table 19.46 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Mathematics teacher behaviour (Module 16)*

#### **Cognitive activation in mathematics: Foster reasoning (COGACRCO)**

Students’ frequency ratings of how often their mathematics teacher showed a range of behaviours indicative of fostering mathematics reasoning during the ongoing school year (e.g., “The teacher asked us to explain our reasoning when solving a mathematics problem.”, “The teacher asked us to defend our answer to a mathematics problem.”) in question ST285 were scaled into the index of “Cognitive activation in mathematics: Foster reasoning”. Note that this scale used a within-construct matrix sampling design. Each of the nine items included in this scale had five response options (“Never or almost never”, “Less than half of the lessons”, “About half of the lessons”, “More than half of the lessons”, “Every lesson or almost every lesson”). Table 19.47 shows the item wording and item parameters for the items in this scale.

#### **Cognitive activation in mathematics: Encourage mathematical thinking (COGACMCO)**

Students’ frequency ratings of how often their mathematics teacher showed a range of behaviours indicative of encouraging mathematical thinking during the ongoing school year (e.g., “The teacher encouraged us to “think mathematically.”, “The teacher asked us how different topics are connected to a bigger mathematical idea.”) in question ST283 were scaled into the index of “Cognitive activation in mathematics: Encourage mathematical thinking”. Note that this scale used a within-construct matrix sampling design. Each of the nine items included in this scale had five response options (“Never or almost never”, “Less than half of the lessons”, “About half of the lessons”, “More than half of the lessons”, “Every lesson or almost every lesson”). Table 19.48 shows the item wording and item parameters for the items in this scale.

### **Disciplinary climate in mathematics (DISCLIM)**

Students' frequency ratings of how often a range of situations occurred in their mathematics lessons (e.g., "Students do not listen to what the teacher said.", "Students get distracted by using <digital resources> (e.g., smartphones, websites, apps).") in question ST273 were scaled into the index of "Disciplinary climate in mathematics". Note that this scale was linked to the DISCLIMA scale in PISA 2012 and was scaled using the PCM, in line with the model used in PISA 2012. Also, this scale used a within-construct matrix sampling design. Each of the seven items included in this scale had four response options ("Every lesson", "Most lessons", "Some lessons", "Never or almost never"). Table 19.49 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

#### *Parental/guardian involvement and support (Module 19)*

### **Family support (FAMSUP)**

Students' ratings of how often their parents or someone else in their family engaged in a range of behaviours indicative of family support (e.g., "Discuss how well you are doing at school", "Spend time just talking with you") in question ST300 were scaled into the index of "Family support". Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had five response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day"). Table 19.50 shows the item wording and item parameters for the items in this scale.

#### *Creative thinking (Module 20)*

### **Creative peers and family environment (CREATFAM)**

Students' ratings of their agreement with statements about the degree to which creative thinking is fostered and supported by their peers and family environment (e.g., "My friends are open to new ideas.", "At home, I am encouraged to use my imagination.") in question ST336 were scaled into the index of "Creative peers and family environment". Note that this scale used a within-construct matrix sampling design. Each of the six items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.51 shows the item wording and item parameters for the items in this scale.

### **Creative school and class environment (CREATSCH)**

Students' ratings of their agreement with statements about the degree to which creative thinking is fostered and supported in their school and class environment (e.g., "My teachers value students' creativity.", "At school, I am given a chance to express my ideas.") in question ST335 were scaled into the index of "Creative school and class environment". Note that this scale used a within-construct matrix sampling design. Each of the six items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.52 shows the item wording and item parameters for the items in this scale.

### **Creative thinking self-efficacy (CREATEFF)**

Students' ratings of how confident they felt about having to do a range of tasks reflective of creative thinking skills (e.g., "Coming up with creative ideas for school projects", "Inventing new things") in question ST334 were scaled into the index of "Creative thinking self-efficacy". Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had four response options ("Not at all confident", "Not very confident", "Confident", "Very confident"). Table 19.53 shows the item wording and item parameters for the items in this scale.

### **Creativity and openness to intellect (CREATOP)**

Students' ratings of their agreement with statements regarding their own views on their creativity and openness to intellect (e.g., "Doing something creative satisfies me.", "I like games that challenge my creativity.") in question ST340 were scaled into the index of "Creativity and openness to intellect". Note that this scale used a within-construct matrix sampling design. Each of the 10 items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.54 shows the item wording and item parameters for the items in this scale.

### **Imagination and adventurousness (IMAGINE)**

Students' ratings of their agreement with statements regarding their own views on their imagination and adventurousness (e.g., "I have difficulty using my imagination.", "Coming up with new ideas is satisfying to me.") in question ST342 were scaled into the index of "Imagination and adventurousness". Note that this scale used a within-construct matrix sampling design. Each of the seven items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.55 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### **Openness to art and reflection (OPENART)**

Students' ratings of their agreement with statements regarding their own views on their openness to art and reflection (e.g., "I enjoy creating art.", "I reflect on movies I watch.") in question ST341 were scaled into the index of "Openness to art and reflection". Each of the five items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.56 shows the item wording and item parameters for the items in this scale.

### **Participation in creative activities at school (CREATAS)**

Students' ratings of how often they participated in creative activities that were available in their school (e.g., "Art classes/activities (e.g., painting, drawing)", "Debate club") in question ST337 were scaled into the index of "Participation in creative activities at school". Note that the activities sampled in this question are the same as the activities in the "outside of school" version of this question (CREATOOS – ST338). Each of the eight items included in this scale had five substantive response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day") and an additional response option "Not available at school" which was recoded as missing prior to scaling. Table 19.57 shows the item wording and item parameters for the items in this scale. It also indicates how the response categories were recoded prior to scaling.

### **Participation in creative activities outside of school (CREATOOS)**

Students' ratings of how often they participated in creative activities outside of school (e.g., "Art classes/activities (e.g., painting, drawing)", "Debate club") in question ST338 were scaled into the index of "Participation in creative activities outside of school". Note that the activities sampled in this question are the same as the activities in the "at school" version of this question (CREATAS – ST337). Each of the eight items included in this scale had five substantive response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day") and an additional response option "Not available" which was recoded as missing prior to scaling. Table 19.58 shows the item wording and item parameters for the items in this scale. It also indicates how the response categories were recoded prior to scaling.

*Global crises (Module 21)*

Note that the questions in this module were skipped for students who reported that their school had not been closed for more than a week due to COVID-19 in question ST347.

**Family support for self-directed learning (FAMSUPSL)**

Students' frequency ratings of how often someone in their family provided specific kinds of learning support (e.g., "Help me create a learning schedule"; "Help me access learning materials online") while the school building was closed due to COVID-19 in question ST353 were scaled into the index of "Family support for self-directed learning". Note that this scale used a within-construct matrix sampling design. Each of the eight items included in this scale had four response options ("Never", "A few times", "About once or twice a week", "Every day or almost every day"). Table 19.59 shows the item wording and item parameters for the items in this scale.

**Feelings about learning at home (FEELLAH)**

Students' ratings of their agreement with statements about how they felt about learning at home (e.g., "I enjoyed learning by myself.", "My teachers were well prepared to provide instruction remotely.") while the school building was closed due to COVID-19 in question ST354 were scaled into the index of "Feelings about learning at home". Note that this scale used a within-construct matrix sampling design. Each of the six items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.60 shows the item wording and item parameters for the items in this scale.

**Problems with self-directed learning (PROBSELF)**

Students' frequency ratings of how often they had various problems completing their school work (e.g., "Problems with Internet access", "Problems with understanding my school assignments") while their school building was closed due to COVID-19 in question ST352 were scaled into the index of "Problems with self-directed learning". Note that this scale used a within-construct matrix sampling design. Each of the eight items included in this scale had four response options ("Never", "A few times", "About once or twice a week", "Every day or almost every day"). Table 19.61 shows the item wording and item parameters for the items in this scale.

**Self-directed learning self-efficacy (SDLEFF)**

Students' ratings of how confident they felt about having to do a range of self-directed learning tasks (e.g., "Finding learning resources online on my own", "Completing school work independently") should their school building close again in the future in question ST355 were scaled into the index of "Self-directed learning self-efficacy". Note that this scale used a within-construct matrix sampling design. Each of the eight items included in this scale had four response options ("Not at all confident", "Not very confident", "Confident", "Very confident"). Table 19.62 shows the item wording and item parameters for the items in this scale.

**School actions to sustain learning (SCHSUST)**

Students' frequency ratings of how often someone from their school completed an activity to sustain their learning (e.g., "Sent me learning materials to study on my own", "Checked in with me to ensure that I was completing my assignments") while their school building was closed due to COVID-19 in question ST348 were scaled into the index of "School actions to sustain learning". Note that this scale used a within-construct matrix sampling design. Each of the eight items included in this scale had four response options



(“Never”, “A few times”, “About once or twice a week”, “Every day or almost every day”). Table 19.63 shows the item wording and item parameters for the items in this scale.

### Types of learning resources used while school was closed (LEARRES)

Students’ frequency ratings of how often they used specific learning resources (e.g., “Paper textbooks, workbooks, or worksheets”, “Recorded lessons or other digital material provided by teachers from my school”) while the school building was closed due to COVID-19 in question ST351 were scaled into the index of “Types of learning resources used while school was closed”. Note that this scale used a within-construct matrix sampling design. Each of the eight items included in this scale had four response options (“Never”, “A few times”, “About once or twice a week”, “Every day or almost every day”). Table 19.64 shows the item wording and item parameters for the items in this scale.

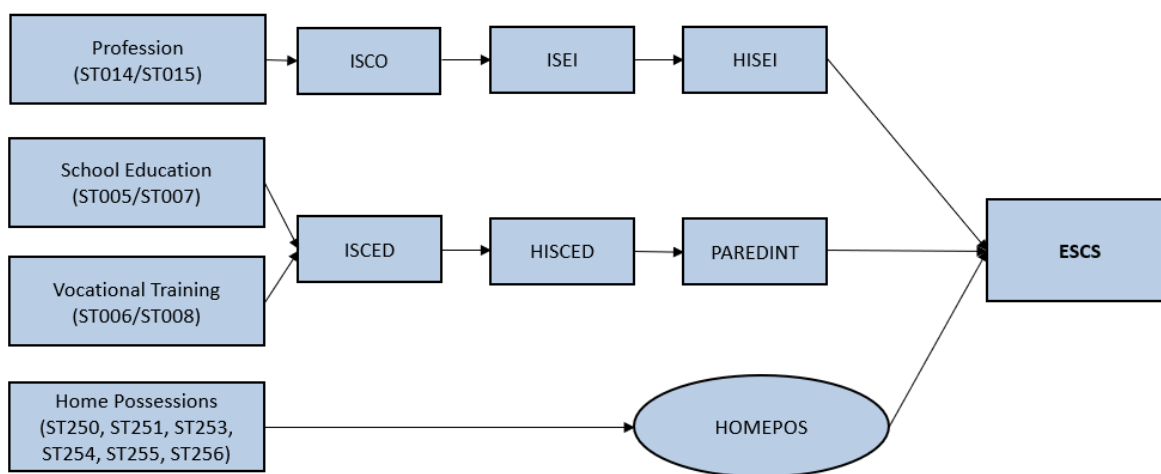
### Complex composite index – Index of economic, social and cultural status (ESCS)

There was only one complex composite index derived from the Student Questionnaire – the index of economic, social and cultural status (ESCS).

#### Components of ESCS

The ESCS score was based on three indicators: highest parental occupation status (HISEI), highest education of parents in years (PAREDINT), and home possessions (HOMEPOS). The rationale for using these three components, which are consistent with the components used in previous PISA cycles, is that socio-economic status is most commonly theoretically conceptualized based on “the big 3” (occupational status, education, and income) (Cowan et al., 2012<sup>[13]</sup>). As no direct income measure is available in the PISA data, the existence of household items has been used as a proxy for family income. Figure 19.4 provides a schematic representation of ESCS and its components.

Figure 19.4. Computation of ESCS in PISA 2022



**HISEI.** For more information on HISEI, refer to the explanation on HISEI in the simple indices section above.

**PAREDINT.** For more information on PAREDINT, refer to the explanation on PAREDINT in the simple indices section above.

**HOMEPOS.** For more information on HOMEPOS, refer to the explanation on HOMEPOS in the IRT scale section above.

### *Computation of ESCS*

The ESCS scores were computed using the same methodology used in PISA 2018 (Avvisati, 2020<sup>[14]</sup>; OECD, 2020<sup>[15]</sup>). For students with missing data on one out of the three components, the missing component was imputed using a regression equation which was created for each country/economy using data from students without any missing components. For each student with a missing component, this regression equation was used to predict the missing component with the two non-missing components and a random value was added to the predicted value to reflect the error of the regression model.<sup>5</sup> If a student had missing data on more than one component, the ESCS score was not computed for the student, and the student's ESCS score was replaced with "99" in the SPSS file and ".M" in the SAS file.

After the imputation process, each of the three components (including the imputed values) was standardised to have a mean of 0 and a standard deviation of 1 across the OECD countries, with each OECD country weighted approximately equally using senate weights.<sup>6</sup> The OECD means and standard deviations that were used to standardise each component of ESCS are displayed in Table 19.65.

Subsequently, the arithmetic mean of the three standardised components was calculated to create a preliminary ESCS score for each student. Lastly, the preliminary ESCS scores were standardised again to have a mean of 0 and a standard deviation of 1 across the OECD countries (again with each country weighted approximately equally using senate weights<sup>7</sup>), producing the final ESCS score for each student. The OECD mean and standard deviation that were used to transform the preliminary ESCS scores into the final ESCS scores are displayed in Table 19.65.

### *ESCS trend scores*

In contrast to the other trend scales in the context questionnaires (for which the scale scores for the current cycle were made to be comparable to the scale scores from a previous cycle), the scores for each component of ESCS and the composite ESCS scores are not comparable to the scores from previous cycles. Instead, each of the component scores for ESCS and the composite ESCS scores for PISA 2012, PISA 2015, and PISA 2018 were recomputed to be comparable to the respective scores for PISA 2022. This was done by recoding the scores for each component of ESCS for PISA 2012, PISA 2015, and PISA 2018 using the coding scheme used in PISA 2022, then recomputing the composite ESCS score for these previous cycles using the ESCS computation methodology used in PISA 2022. More details are provided below.

**HISEI.** Until PISA 2009, ISCO-88 was used to code parental occupation. However, since PISA 2012, parental occupation has been coded using ISCO-08, the most recent version of ISCO. In PISA 2018, the coding scheme for ISEI was updated so that an ISEI value of 17 was attributed to ISCO codes 9701 ("stay-at-home parent"), 9702 ("student"), and 9703 ("social beneficiary"), equivalent to the ISEI value for ISCO code 9000 ("elementary occupations"). This coding scheme was also used in PISA 2022.

To make the HISEI scores for PISA 2012 and PISA 2015 comparable to the HISEI scores for PISA 2018 and PISA 2022, new HISEI scores were created for each student that participated in PISA 2012 and PISA 2015 using the coding scheme used in PISA 2018 and PISA 2022. These new HISEI scores were used in the computation of the trend ESCS scores.

**PAREDINT.** For some countries/economies, the mapping of ISCED levels to years of education was updated in 2009, 2015, and 2018, taking into account changes in the countries/economies' educational systems. In PISA 2022, PAREDINT was updated again to map each ISCED level (based on ISCED-11) to the PISA 2018 cumulative years of education values, as presented in Table 19.7.

To make the PAREDINT scores for PISA 2012, PISA 2015, and PISA 2018 comparable to the PAREDINT scores for PISA 2022, new PAREDINT scores were created for each student that participated in the previous cycles the mapping presented in Table 19.7. These new PAREDINT scores were used in the computation of the trend ESCS scores.

**HOMEPOS.** Indicators of HOMEPOS have been dropped or added in all PISA cycles, taking into account the social, technical, and economic changes in the participating countries/economies. Moreover, the method for estimating HOMEPOS changed in PISA 2009, PISA 2012, and PISA 2015.

To make the HOMEPOS scores comparable across cycles, prior to scaling, the response categories for some items in PISA 2012, PISA 2015, and PISA 2018 were collapsed to align with the response categories used in PISA 2022 (as presented in Table 19.16). Then, the HOMEPOS WLEs for each student that participated in the past three cycles were re-estimated by fixing the item parameters to the parameters that were estimated for each group in PISA 2022 (either the international parameters or the group's unique parameters, depending on whether the item parameters were released for the group in PISA 2022). For items that were not administered in PISA 2022, new international parameters were estimated by pooling data across all cycles and groups in which the item had been administered, then the item parameters were released until all groups in all cycles had an RMSD under 0.25. As an exception, unique parameters were estimated for all country/economy-specific items for all groups and all cycles. These newly estimated HOMEPOS WLEs were used in the computation of the trend ESCS scores.

**ESCS.** Prior to PISA 2018, the ESCS scores were computed using a principal component analysis (PCA), although there were differences across cycles regarding which countries/economies were included in the PCA and how the scores were standardised. In PISA 2018, the ESCS scores were computed as the arithmetic mean of the three components, with each of the component scores and the composite ESCS score standardised to have a mean of 0 and standard deviation of 1 across the OECD countries (with each country weighted approximately equally using senate weights<sup>8</sup>). As noted above, this methodology was also used in PISA 2022.

To make the ESCS scores for PISA 2012, PISA 2015, and PISA 2018 comparable to the ESCS scores for PISA 2022, new ESCS scores were computed for the previous cycles using the methodology used in PISA 2022. Specifically, for students with missing data on one out of the three components, the missing component was imputed using a regression equation which was created for each country/economy in each cycle using data from students without any missing components. For each student with a missing component, this regression equation was used to predict the missing component with the two non-missing components and a random value was added to the predicted value to reflect the error of the regression model.<sup>9</sup> If a student had missing data on more than one component, the ESCS score was not computed for the student, and the student's ESCS score was replaced with "99" in the SPSS file and ".M" in the SAS file.

After the imputation process, each of the three components (including the imputed values) was standardised using the OECD mean and standard deviation of the respective component in PISA 2022 (presented in Table 19.65 above). Next, the arithmetic mean of the three standardised components was calculated to create a preliminary ESCS score for each student. Lastly, the preliminary ESCS scores were standardised again using the OECD mean and standard deviation of the preliminary ESCS scores in PISA 2022 (also presented in Table 19.65). This process ensured that the trend ESCS scores produced for PISA 2012, PISA 2015, and PISA 2018 were directly comparable to the ESCS scores produced for PISA 2022.

## Financial Literacy Questionnaire derived variables

The Financial Literacy Questionnaire is an international option that countries/economies could choose to implement. It was administered to students after they had completed the Student Questionnaire. It

addresses familiarity of students related to financial literacy and their confidence about financial matters. There were 10 variables derived from this questionnaire, including one simple DV and nine IRT scaled DVs. An overview of all DVs in this questionnaire is shown in Table 19.66 and each are described in the following sections.

### **Simple questionnaire indices**

#### *Familiarity with concepts of finance (FCFMLRTY)*

Students' ratings of how familiar they were with various financial topics in question FL164 were used to derive an indicator of familiarity with concepts of finance. There were three response options ("Never heard of it", "Heard of it, but I don't recall the meaning", "Learnt about it, and I know what I means"). For each item, a value of "1" was assigned to "Learnt about it, and I know what I means" responses, and all other responses were assigned a value of "0". This index was constructed as the sum of values across all 16 items. Values range from "0" to "16".

### **Derived variables based on IRT scaling**

The Financial Literacy Questionnaire provided data for nine DVs based on IRT scaling. The Cronbach's alpha for each scale and group are presented in Table 19.67, the number of items with international parameters for each scale and group are presented in Table 19.68, the number of trend items with international parameters for each trend scale and group are presented in Table 19.69, and the groups that did not administer each scale are presented in Table 19.70 (in this case, the scale scores for the individuals in the group were replaced with "99" in the SPSS file and ".M" in the SAS file). Note that there were no countries/economies for which the scale scores were suppressed for the scales in the Financial Literacy Questionnaire.

### **Financial education in school lessons (FLSCHOOL)**

Students' frequency ratings of how often they encountered financial tasks and activities in school lessons (e.g., "Describing the purpose and uses of money", "Exploring ways of planning to pay an expense") in question FL166 were scaled into the index of "Financial education in school lessons". Note that this scale was linked to the FLSCHOOL scale in PISA 2018. Each of the six items included in this scale had three response options ("Never", "Sometimes", "Often"). Table 19.71 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### **Financial education in school lessons – Multiple subjects (FLMULTSB)**

Students' responses to questions about where they encountered lessons about financial topics (e.g., "During your mathematics class", "During classes about economics or business") in question FL174 were scaled into the index of "Financial education in school lessons – Multiple Subjects (FLMULTSB)". Each of the seven items included in this scale had two substantive response options ("Yes", "No") and two additional response options ("I don't know.", "I don't have this class.") which were recoded as missing prior to scaling. Table 19.72 shows the item wording and item parameters for the items in this scale. It also indicates how the response categories were recoded prior to scaling.

### **Parental involvement in matters of financial literacy (FLFAMILY)**

Students' frequency ratings of how often they discuss various financial issues with their parents (e.g., "Your spending decisions", "Shopping online") in question FL167 were scaled into the index of "Parental involvement in matters of financial literacy". Note that this scale was linked to the FLFAMILY scale in PISA

2018. Each of the seven items included in this scale had four response options (“Never or hardly ever”, “Once or twice a month”, “Once or twice a week”, “Almost every day”). Table 19.73 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### ***Access to money and financial projects – Sources of money (ACCESSFP)***

Students’ frequency ratings about how often their money came from different sources (e.g., “An allowance or pocket money for doing chores at home”, “Working in a family business”) in question FL170 were scaled into the index of “Access to money and financial projects – Sources of Money (ACCESSFB)”. Each of the seven items included in this scale had five response options (“Never or almost never”, “About once or twice a year”, “About once or twice a month”, “About once or twice a week”, “Every day or almost every day”). Table 19.74 shows the item wording and item parameters for the items in this scale.

### ***Confidence about financial matters (FLCONFIN)***

Students’ ratings of their confidence with various financial matters (e.g., “Understanding bank statements”, “Keeping track of my account balance”) in question FL162 were scaled into the index of “Confidence about financial matters”. Note that this scale was linked to the FLCONFIN scale in PISA 2018. Each of the six items included in this scale had four response options (“Not at all confident”, “Not very confident”, “Confident”, “Very confident”). Table 19.75 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### ***Confidence about financial matters using digital devices (FLCONICT)***

Students’ ratings of their confidence in doing various financial tasks with electronic devices (e.g., “Transferring money”, “Paying with a mobile device (e.g., mobile phone or tablet) instead of using cash”) in question FL163 were scaled into the index of “Confidence about financial matters using digital devices”. Note that this scale was linked to the FLCONICT scale in PISA 2018. Each of the five items included in this scale had four response options (“Not at all confident”, “Not very confident”, “Confident”, “Very confident”). Table 19.76 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### ***Access to money and financial products – Financial activities (ACCESSFA)***

Students’ frequency ratings of how often they completed different financial activities (e.g., “Checked how much money you have”, “Saved money at home”) in question FL171 were scaled into the index of “Access to money and financial products – Financial activities (ACCESSFA)”. Each of the 11 items included in this scale had five response options (“Never or almost never”, “About once or twice a year”, “About once or twice a month”, “About once or twice a week”, “Every day or almost every day”). Table 19.77 shows the item wording and item parameters for the items in this scale.

### ***Attitudes towards and confidence about financial matters (ATTCONFM)***

Students’ rating of their agreement with different statements about their attitudes towards and confidence about financial matters (e.g., “I enjoy talking about money matters.”, “I know how to manage my money.”) in question FL169 were scaled into the index of “Attitudes towards and confidence about financial matters (ATTCONFM)”. Each of the seven items included in this scale had four response options (“Strongly disagree”, “Disagree”, “Agree”, “Strongly agree”). Table 19.78 shows the item wording and item parameters for the items in this scale.

### ***Friends' influence on financial matters (FRINFLFM)***

Students' ratings of their agreement with various statements about their friends' influence on finance decisions (e.g., "My friends have a strong influence on my spending decisions.", "Sometimes I spend more than I would like when I am with my friends.") in question FL172 were scaled into the index of "Friends' influence on financial matters". Each of the four items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.79 shows the item wording and item parameters for the items in this scale.

### **ICT Familiarity Questionnaire derived variables**

The ICT Familiarity Questionnaire is an international option that countries/economies could choose to implement. It was administered to students after they had completed the Student Questionnaire. There were 15 variables derived from this questionnaire, including three simple DVs and 12 IRT scaled DVs. All of the IRT scaled DVs were new for PISA 2022, as the ICT framework had been revised for this cycle. An overview of all DVs in this questionnaire is shown in Table 19.80 and each are described in the following sections.

#### ***Simple questionnaire indices***

##### *Availability and usage of ICT at school (ICTAVSCH)*

The availability of ICT at school was gathered from IC170 where students' frequency ratings of how often they use various digital resources at school (e.g., "Desktop or laptop computer", "Smartphone") was used for the index of "ICT availability at school". Each of the seven items in this question included six response options ("Never or almost never", "About once or twice a month", "About once or twice a week", "Every day or almost every day", "Several times a day", "This resource is not available to me at school"). The index was calculated as the number of all seven items that were marked with a value other than "This resource is not available to me at school", thus ranging from 0-7. Items 2-4 were included in various previous versions of the ICT Questionnaire.

##### *Availability and usage of ICT outside of school (ICTAVHOM)*

The availability of ICT outside of school was gathered from IC171 where students' frequency ratings of how often they use various digital resources outside of school (e.g., "Desktop or laptop computer", "Smartphone") was used for the index of "ICT use outside of school". Each of the six items in this question included six response options ("Never or almost never", "About once or twice a month", "About once or twice a week", "Every day or almost every day", "Several times a day", "This resource is not available to me outside of school"). For each of the six items, a score of "0" was assigned when students choose the "This resource is not available to me outside of school" response options and all other responses were coded "1". The index was calculated as the sum of "0" and "1" designations across the six items that were marked with a value other than "This resource is not available to me at school", thus ranging from 0-6. Items 2-4 were included in various previous versions of the ICT Questionnaire.

##### *Distress from online content and cyberbullying (ICTDISTR)*

Students' ratings of how upset they were when various situation occurred online (e.g., "Encountering content online that was inappropriate for my age", "Receiving unkind, vulgar or offending messages, comments or videos") in question IC181 were scaled into the index of "Distress from online content and cyberbullying". Each item included five response options ("This did not happen to me", "Not at all upset",

“A little upset”, “Quite upset”, “Very upset”). Values in this index range from 0-16, with “This did not happen to me” recoded as a missing variable, “Not at all upset” coded “1”, “A little upset” coded “2”, “Quite upset” coded “3”, and “Very upset” coded “4”. Values across all items were summed.

### ***Derived variables based on IRT scaling***

The ICT Familiarity Questionnaire provided data for 12 DVs based on IRT scaling. The Cronbach’s alpha for each scale and group are presented in Table 19.81, the number of items with international parameters for each scale and group are presented in Table 19.82, the countries/economies for which the scale scores were suppressed for each scale are presented in Table 19.83 (in this case, the scale scores for the individuals in the country/economy were replaced with “97” in the SPSS file and “.N” in the SAS file), and the groups that did not administer each scale are presented in Table 19.84 (in this case, the scale scores for the individuals in the group were replaced with “99” in the SPSS file and “.M” in the SAS file).

#### *ICT availability at school (ICTSCH)*

Students’ frequency ratings of how often they use various digital resources at school (e.g., “Desktop or laptop computer”, “Smartphone (i.e., mobile phone with internet access)”) in question IC170 were scaled into the index of “ICT availability at school”. Each of the seven items included in this scale had six response options (“Never or almost never”, “About once or twice a month”, “About once or twice a week”, “Every day or almost every day”, “Several times a day”, “This resource is not available to me at school”). “This resource is not available to me at school” was recoded as 0, while the five other response options were recoded as 1 prior to scaling. Table 19.85 shows the item wording and item parameters for the items in this scale.

#### *ICT availability outside school (ICTHOME)*

Students’ frequency ratings of how often they use various digital resources outside of school (e.g., “Desktop or laptop computer”, “Smartphone (i.e., mobile phone with internet access)”) in question IC171 were scaled into the index of “ICT availability outside school”. Each of the six items included in this scale had six response options (“Never or almost never”, “About once or twice a month”, “About once or twice a week”, “Every day or almost every day”, “Several times a day”, “This resource is not available to me outside of school”). “This resource is not available to me outside of school” was recoded as 0, while the five other response options were recoded as 1 prior to scaling. Table 19.86 shows the item wording and item parameters for the items in this scale.

#### *Quality of access to ICT (ICTQUAL)*

Students’ ratings of their agreement with various statements about ICT resources at their school (e.g., “There are enough digital devices with access to the Internet at my school.”, “The school’s Internet speed is sufficient.”) in question IC172 were scaled into the index of “Quality of access to ICT”. Each of the nine items included in this scale had four response options (“Strongly disagree”, “Disagree”, “Agree”, “Strongly agree”). Table 19.87 shows the item wording and item parameters for the items in this scale.

#### *Subject-related ICT use during lessons (ICTSUBJ)*

Students’ frequency ratings of how often digital resources are used in various subject lessons (e.g., “Mathematics”, “Science”) in question IC173 were scaled into the index of “Subject-related ICT use during lessons”. Each of the four items included in this scale had five substantive response options (“Never or almost never”, “In less than half of the lessons”, “In about half of the lessons”, “In more than half of the lessons”, “In every or almost every lesson”) and an additional response option “I do not have this subject” which was recoded as missing prior to scaling. Table 19.88 shows the item wording and item parameters for the items in this scale.

### *Use of ICT in enquiry-based learning activities (ICTENQ)*

Students' frequency ratings of how often they use digital resources for various school-related activities (e.g., "Create a multi-media presentation with pictures, sound or video", "Track the progress of your own work or projects") in question IC174 were scaled into the index of "Use of ICT in enquiry-based learning activities". Each of the 10 items included in this scale had five response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day"). Table 19.89 shows the item wording and item parameters for the items in this scale.

### *Support or feedback via ICT (ICTFEED)*

Students' frequency ratings of how often they use digital resources in various activities related to support or feedback (e.g., "Read or listen to feedback sent by my teachers regarding my work and academic results", "Read or listen to feedback sent by other students on my work") in question IC175 were scaled into the index of "Support or feedback via ICT". Each of the four items included in this scale had five response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day"). Table 19.90 shows the item wording and item parameters for the items in this scale.

### *Use of ICT for school activities outside of the classroom (ICTOUT)*

Students' frequency ratings of how often they use digital resources for various school-related activities outside of the classroom (e.g., "See my grades or results from specific assignments (e.g., homework or tests)", "Communicate with my teacher") in question IC176 were scaled into the index of "Use of ICT for school activities outside of the classroom". Each of the eight items included in this scale had five response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day"). Table 19.91 shows the item wording and item parameters for the items in this scale.

### *Frequency of ICT activity – Weekday (ICTWKDY)*

Students' frequency ratings of how often they did various leisure activities using ICT during a typical week day (e.g., "Play video-games (using my smartphone, a gaming console or an online platform or apps)", "Look for practical information online (e.g., find a place, book a train ticket, buy a product)") in question IC177 were scaled into the index of "Frequency of ICT activity – Weekday". Each of the seven items included in this scale had six response options ("No time at all", "Less than 1 hour a day", "Between 1 and 3 hours a day", "More than 3 hours and up to 5 hours a day", "More than 5 hours and up to 7 hours a day", "More than 7 hours a day"). Table 19.92 shows the item wording and item parameters for the items in this scale.

### *Frequency of ICT activity – Weekend (ICTWKEND)*

Students' frequency ratings of how often they did various leisure activities using ICT during a typical weekend day (e.g., "Play video-games (using my smartphone, a gaming console or an online platform or apps)", "Look for practical information online (e.g., find a place, book a train ticket, buy a product)") in question IC178 were scaled into the index of "Frequency of ICT activity – Weekend". Each of the seven items included in this scale had six response options ("No time at all", "Less than 1 hour a day", "Between 1 and 3 hours a day", "More than 3 hours and up to 5 hours a day", "More than 5 hours and up to 7 hours a day", "More than 7 hours a day"). Table 19.93 shows the item wording and item parameters for the items in this scale.



### *Views of regulated ICT use in school (ICTREG)*

Students' ratings of their agreement with various statements about regulation of ICT use at school (e.g., "Students should not be allowed to bring mobile phones to class.", "The school should set up filters to prevent students from playing games online.") in question IC179 were scaled into the index of "Views of regulated ICT use in school". Each of the six items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.94 shows the item wording and item parameters for the items in this scale.

### *Students' practices regarding online information (ICTINFO)*

Students' ratings of their agreement with various statements about their practices regarding online information (e.g., "When searching for information online I compare different sources.", "I discuss the accuracy of online information with friends or other students.") in question IC180 were scaled into the index of "Students' practices regarding online information". Each of the six items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.95 shows the item wording and item parameters for the items in this scale.

### *Self-efficacy in digital competencies (ICTEFFIC)*

Students' ratings of how well they can do various tasks using digital resources (e.g., "Search for and find relevant information online", "Write or edit text for a school assignment") in question IC183 were scaled into the index of "Self-efficacy in digital competencies". Each of the 14 items included in this scale had four substantive response options ("I cannot do this", "I struggle to do this on my own", "I can do with a bit of effort", "I can easily do this") and an additional response option "I don't know what this is" which was recoded as missing prior to scaling. Table 19.96 shows the item wording and item parameters for the items in this scale.

## **Well-Being Questionnaire derived variables**

The Well-Being Questionnaire is an international option that countries/economies could choose to implement. It was administered to students after they had completed the Student Questionnaire. It addresses the well-being of students. There were seven variables derived from this questionnaire, including one simple DV and six IRT scaled DVs. An overview of all DVs in this questionnaire is shown in Table 19.97 and each are described in the following sections.

### **Simple questionnaire indices**

#### *Body mass index (STUBMI)*

The only simple DV from the Well-Being Questionnaire is STUBMI, indicating the student's body mass index (STUBMI). It is based on two questions, WB151 and WB152, which asked about the weight and the height of the student, respectively, in the units of measurement that are more common in the respective country/economy. The index is constructed as it was in PISA 2018. Specifically, the index was constructed as the weight (transformed to kilograms) divided by the square of the body height (transformed to metres).

### **Derived variables based on IRT scaling**

The Well-Being Questionnaire provided data for six DVs based on IRT scaling. The Cronbach's alpha for each scale and group are presented in Table 19.98, the number of items with international parameters for each scale and group are presented in Table 19.99, the number of trend items with international

parameters for each trend scale and group are presented in Table 19.100, and the groups that did not administer each scale are presented in Table 19.101 (in this case, the scale scores for the individuals in the group were replaced with “99” in the SPSS file and “.M” in the SAS file). Note that there were no countries/economies for which the scale scores were suppressed for the scales in the Well-Being Questionnaire.

#### *Body image (BODYIMA)*

Students’ ratings of their agreement with statements about their body image (e.g., “I like my look just the way it is.”, “I like my body.”) in question WB153 were scaled into the index of “Body image”. Note that this scale was linked to the BODYIMA scale in PISA 2018. Each of the five items included in this scale had four substantive response options (“Strongly disagree”, “Disagree”, “Agree”, “Strongly agree”) and an additional response option “I don’t have an opinion” which was recoded as missing prior to scaling. Table 19.102 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

#### *Social connection to parents (SOCONPA)*

Students’ ratings of how often their parents engage in various activities (e.g., “Show that they care”, “Encourage me to make my own decisions”) in question WB163 were scaled into the index of “Social connection to parents”. Note that this scale was linked to the SOCONPA scale in PISA 2018. Each of the six items included in this scale had three response options (“Almost never”, “Sometimes”, “Almost always”). Table 19.103 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

#### *Students’ life satisfaction across domains (LIFESAT)*

Students’ ratings of their satisfaction with different areas of their lives (e.g., “Your health”, “The neighbourhood you live in”) in question WB155 were scaled into the index of “Students’ life satisfaction across domains”. Each of the 10 items included in this scale had four response options (“Not at all satisfied”, “Not satisfied”, “Satisfied”, “Totally satisfied”). Table 19.104 shows the item wording and item parameters for the items in this scale.

#### *Psychosomatic symptoms (PSYCHSYM)*

Students’ ratings of how often they experienced different psychosomatic symptoms (e.g., “Headache”, “Stomach pain”) in question WB154 were scaled into the index of “Psychosomatic symptoms”. Each of the nine items included in this scale had five response options (“Rarely or never”, “About every month”, “About every week”, “More than once a week”, “About every day”). Table 19.105 shows the item wording and item parameters for the items in this scale.

#### *Social connections: Ease of communication about worries and concerns (SOCCON)*

Students’ ratings of how easy it is to communicate about their worries and concerns with different people (e.g., “Your father”, “Your brother(s)”) in question WB162 were scaled into the index of “Social connections: Ease of communication about worries and concerns”. Each of the nine items included in this scale had four substantive response options (“Very difficult”, “Difficult”, “Easy”, “Very Easy”) and an additional response option “I don’t have or see this person” which was recoded as missing prior to scaling. Table 19.106 shows the item wording and item parameters for the items in this scale.

### *Experienced well-being – Previous day (EXPWB)*

Students' responses regarding their experienced well-being in the previous day (e.g., “Were you treated with respect all day yesterday?”, “Did you smile or laugh a lot yesterday?”) in question WB178 were scaled into the index of “Experienced well-being – Previous day”. Each of the six items included in this scale had two response options (“Yes”, “No”). Note that prior to scaling, all responses were recoded as missing if the student did not respond “yes” to WB178Q07JA (“Was yesterday a typical day?”). Table 19.107 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

## **Parent Questionnaire derived variables**

The Parent Questionnaire is an international option that countries/economies could choose to implement. It was administered to the parents of students participating in the PISA assessment. There were 12 variables derived from this questionnaire, including one simple DV and 11 IRT scaled DVs. An overview of all DVs in this questionnaire is shown in Table 19.108 and each are described in the following sections.

### **Simple questionnaire indices**

#### *Parents' expectations in child's future educational career (PAREXPT)*

Parents' responses to a list of possible educational levels they expected their children to complete in question PA183 were transformed into the index of “Parents' expectations in child's future educational career”. The categories were specified using country-specific terms that were understood by the respondents. Each qualification was mapped to the ISCED classification of educational levels [see *ISCED 2011 Operational Manual: Guidelines for Classifying National Educational Programmes and Related Qualifications* (OECD/Eurostat/UNESCO Institute for Statistics, 2015<sub>[16]</sub>)]. Values on the index ranged from “Less than ISCED level 2” to “ISCED level 8” and scores were assigned as noted in Table 19.109.

### **Derived variables based on IRT scaling**

The Parent Questionnaire provided data for 11 DVs based on IRT scaling. The Cronbach's alpha for each scale and group are presented in Table 19.110, the number of items with international parameters for each scale and group are presented in Table 19.111, the number of trend items with international parameters for each trend scale and group are presented in Table 19.112, the countries/economies for which the scale scores were suppressed for each scale are presented in Table 19.113 (in this case, the scale scores for the individuals in the country/economy were replaced with “97” in the SPSS file and “.N” in the SAS file), and the groups that did not administer each scale are presented in Table 19.114 (in this case, the scale scores for the individuals in the group were replaced with “99” in the SPSS file and “.M” in the SAS file).

#### *Current parental/guardian support (CURSUPP)*

Parents' frequency ratings of how often they or someone else in their home provides education-related support (e.g., “Discuss how well my child is doing at school,” “Talk to my child about any problems he/she may have at school”) in question PA003 were scaled into the index of “Current parental/guardian support”. Note that this scale was linked to the CURSUPP scale in PISA 2018. Each of the 14 items included in this scale had five response options (“Never or hardly ever”, “Once or twice a year”, “Once or twice a month”, “Once or twice a week”, “Every day or almost every day”). Table 19.115 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Parent attitudes toward mathematics (PQMIMP)*

Parents' ratings of their agreement with statements about the importance of mathematical knowledge (e.g., "Most jobs today require some mathematics knowledge and skills.", "It is an advantage in the job market to have good mathematics knowledge and skills.") in question PA196 were scaled into the index of "Parent attitudes toward mathematics". Note that this scale was linked to the PQMIMP scale in PISA 2012 and was scaled using the PCM, in line with the model used in PISA 2012. Each of the four items included in this scale had four response options ("Strongly agree", "Agree", "Disagree", "Strongly disagree"). Table 19.116 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items and which items were reverse-coded prior to scaling.

### *Mathematics career (PQMCAR)*

Parents' responses to questions about mathematics-related careers (e.g., "Does anybody in your family (including you) work in a <mathematics-related career>?", "Does your child show an interest in working in a <mathematics-related career>?") in question PA197 were scaled into the index of "Mathematics career". Note that this scale was linked to the PQMCAR scale in PISA 2012 and was scaled using the PCM, in line with the model used in PISA 2012. Each of the five items included in this scale had two response options ("Yes", "No"). Table 19.117 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items and which items were reverse-coded prior to scaling.

### *Parental involvement (PARINVOL)*

Parents' responses to questions about their involvement in their child's schooling in the past year (e.g., "Discussed my child's progress with a teacher on my own initiative", "Attended a scheduled meeting or conferences for parents") in question PA008 were scaled into the index of "Parental involvement". Each of the 10 items included in this scale had two substantive response options ("Yes", "No") and an additional response option "Not supported by school" which was recoded as missing prior to scaling. Table 19.118 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *School quality (PQSCHOOL)*

Parents' ratings of their agreement with statements about school quality (e.g., "Most of my child's school teachers seem competent and dedicated.", "Standards of achievement are high in my child's school.") in question PA007 were scaled into the index of "School quality". Note that this scale was linked to the PQSCHOOL scale in PISA 2018. Each of the seven items included in this scale had four response options ("Strongly agree", "Agree", "Disagree", "Strongly disagree"). Table 19.119 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling and which items are trend items.

### *School policies for parental involvement (PASCHPOL)*

Parents' ratings of their agreement with statements about school policies for parental involvement (e.g., "My child's school provides effective communication between the school and families.", "My child's school involves parents in the school's decision-making process.") in question PA007 were scaled into the index of "School policies for parental involvement". Note that this scale was linked to the PASCHPOL scale in PISA 2018. Each of the six items included in this scale had four response options ("Strongly agree", "Agree", "Disagree", "Strongly disagree"). Table 19.120 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling and which items are trend items.

### *Parents' attitudes towards immigrants (ATTIMMP)*

Parents' ratings of their agreement with statements about immigrants (e.g., "Immigrant children should have the same opportunities for education that other children in the country have.", "Immigrants who live in a country for several years should have the opportunity to vote in elections.") in question PA167 were scaled into the index of "Parents' attitudes towards immigrants". Note that this scale was linked to the ATTIMMP scale in PISA 2018. Each of the four items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.121 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Creative home environment (CREATHME)*

Parents' ratings of their agreement with statements about creativity in the home environment (e.g., "In our family, we encourage participating in extra-curricular activities that require creativity.", "At home, we try to fix things that are broken.") in question PA185 were scaled into the index of "Creative home environment". Each of the nine items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.122 shows the item wording and item parameters for the items in this scale.

### *Participation in creative activities outside of school (CREATACT)*

Parents' ratings of how often their child participated in creative activities outside of school (e.g., "Art classes/activities (e.g., painting, drawing)", "Debate club") in question PA186 were scaled into the index of "Participation in creative activities outside of school". Each of the eight items included in this scale had five substantive response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day") and an additional response option "Not available" which was recoded as missing prior to scaling. Table 19.123 shows the item wording and item parameters for the items in this scale.

### *Creativity and openness to intellect (CREATOPN)*

Parents' ratings of their agreement with statements regarding their views on their own creativity and openness to intellect (e.g., "I am very creative.", "I enjoy projects that require creative solutions.") in question PA188 were scaled into the index of "Creativity and openness to intellect". Each of the nine items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.124 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Openness to creativity: Other's report (CREATOR)*

Parents' ratings of their agreement with statements regarding their views about their child's creativity (e.g., "My child is very creative.", "My child enjoys projects that require creative solutions.") in question PA189 were scaled into the index of "Openness to creativity: Other's report". Each of the eight items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.125 shows the item wording and item parameters for the items in this scale.

## **School Questionnaire derived variables**

The School Questionnaire consisted mainly of questions used in previous cycles. There were 57 variables derived from this questionnaire, including 32 simple DVs and 25 IRT scaled DVs. An overview of all DVs in this questionnaire is shown in Table 19.126 and each are described in the following sections. The simple

and scaled DVs are organised first by framework module (please see Chapter 5 for a description of modules) and then alphabetical order within modules.

### **Simple questionnaire indices**

#### *Out-of-school experiences (Module 10)*

##### **Creative extra-curricular activities (CREACTIV)**

School principals were asked in SC053 to report what extra-curricular activities their schools offered to 15-year-old students. The two response categories were “Yes” and “No” for the 10 items. The index of creative extra-curricular activities at school (CREACTIV) was computed as the total number of the following 3 activities that occurred at school: i) band, orchestra or choir (SC053Q01TA); ii) school play or school musical (SC053Q02TA); and iii) art club or art activities (SC053Q09TA). The index ranges from 0 to 3. Additionally, a separate DV (SC053D11TA) combines all the customizations across countries to SC053C11TA (please see Annex D).

##### **Mathematics extension courses offered at school (MATHEXC)**

School principals were asked in SC181 to report what additional mathematics lessons are offered at their school. The two response categories were “Yes” and “No”. The index of mathematics extension course offered at school (MATHEXC) was computed from SC181 by assigning schools to one of three different categories based on the type of additional mathematics lessons offered at the school. Schools that responded “Yes” to offering additional mathematics courses without differentiation based on prior achievement (SC181Q03JA) and “No” to offering enrichment (SC181Q01JA) and remedial (SC181Q02JA) mathematics classes were assigned a ‘1’. Schools that responded “Yes” to offering either enrichment mathematics lessons or remedial mathematics lessons were assigned a ‘2’. Schools that responded “Yes” to offering both enrichment and remedial mathematics classes were assigned a ‘3’.

##### **Mathematics-related extra-curricular activities at school (MACTIV)**

School principals were asked in SC053 to report what mathematics-related extra-curricular activities their schools offered to 15-year-old students. The two response categories were “Yes” and “No”. The index of mathematics-related extra-curricular activities at school (MACTIV) was computed as follows. First the question SC181 was assigned the value of ‘1’ if “Yes” was selected for “Enrichment” (SC181Q01JA), “Remedial” (SC181Q02JA), or “Without differentiation depending on the prior achievement level of the students” (SC181Q03JA). SC181 was assigned the value of ‘2’ if “Yes” was selected for both “Enrichment” and “Remedial”. Second, each of three items about a mathematics club (SC053Q05NA), mathematics competitions (SC053Q06NA), or club with a focus on computers (SC053Q08TA) was assigned the value of ‘1’ if a school selected “Yes” to these activities. If a school did not offer one of these three activities (i.e., selected “No”), the corresponding variable received the value of ‘0’. Third, these recoded variables were summed up to result in a range of 0 to 5 for MACTIV. For example, if the purpose of additional lessons was both “Enrichment” and “Remedial” and the school offered a mathematics club, but not mathematics competitions or a club with a focus on computers, the value of MACTIV was coded as “3”.

#### *School type and infrastructure (Module 11)*

##### **Availability of computers (RATCMP1)**

School principals were asked in SC004 to report the number of digital devices available for 15-year-old students at their school. The index of availability of computers (RATCMP1) is the ratio of the number of

desktop or laptop computers available for these students for educational purposes (SC004Q02TA) to the total number of students in the modal grade for 15-year-olds at their school (SC004Q01TA).

#### **Availability of tablet devices (RATTAB)**

School principals were asked in SC004 to report the number of tablet devices or e-book readers available for 15-year-old students at their school for educational purposes (SC004Q08JA). The index of availability of tablet devices (RATTAB) is the ratio of the number of tablet devices available for these students for educational purposes to the total number of students in the modal grade for 15-year-olds at their school (SC004Q01TA).

#### **Computers connected to the Internet (RATCMP2)**

School principals were asked in SC004 to report the number of desktop or laptop computers at their school that are connected to the Internet. The index of computers connected to the Internet (RATCMP2) is the ratio of the number of desktop or laptop computers available for 15-year-olds for educational purposes (SC004Q02TA) to the number of these computers that are connected to the Internet (SC004Q03TA).

#### **Proportion of personnel for pedagogical support (PROPSUPP)**

Principals were asked in SC168 to report the number of personnel for pedagogical support currently working in their school (SC168Q01JA). The proportion of personnel for pedagogical support (PROPSUPP) was calculated by dividing the number of these personnel by the total number of non-teaching staff at the school (TOTSTAFF).

#### **Proportion of school administrative personnel (PROADMIN)**

Principals were asked in SC168 to report the number of school administrative personnel currently working in their school (SC168Q02JA). The proportion of school administrative personnel (PROADMIN) was calculated by dividing the number of these personnel by the total number of non-teaching staff at the school (TOTSTAFF).

#### **Proportion of school management personnel (PROMGMT)**

Principals were asked in SC168 to report the number of school management personnel currently working in their school (SC168Q03JA). The proportion of school management personnel (PROMGMT) was calculated by dividing the number of these personnel by the total number of non-teaching staff at the school (TOTSTAFF).

#### **Proportion of other non-teaching staff (PROOSTAF)**

Principals were asked in SC168 to report the number of other non-teaching staff (i.e., not personnel for pedagogical support, school administrative personnel, or school management personnel) currently working in their school (SC168Q04JA). The proportion of other non-teaching staff (PROOSTAF) was calculated by dividing the number of these personnel by the total number of non-teaching staff at the school (TOTSTAFF).

#### **School size (SCHSIZE)**

The index of school size (SCHSIZE) contains the total enrolment at school. It is based on the enrolment data provided by the school principal in SC002, summing the number of girls (SC002Q02TA) and boys (SC002Q01TA) at a school.

### **School type (SCHLTYPE)**

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 of school type (SCHLTYPE) was constructed by recoding SC013 and SC016. SC013 asks whether the school is public or private, and SC016 asks about the source and proportion of resources (government, resources from students/parents, benefactors/donations, or other). SCHLTYPE has the following three categories:

1. Private independent (if SC013Q01TA=2 and SC016Q01TA < 50), or (SC013Q01TA=2 and  $\text{SUM}(\text{SC016Q02TA}, \text{SC016Q03TA}, \text{SC016Q04TA}) \geq 50$ ).
2. Private Government-dependent (if SC013Q01TA=2 and SC016Q01TA  $\geq 50$ ), or
3. Public (if SC013Q01TA=1).

Since PISA 2018, PRIVATESCH was created from sampling information in order to improve the public/private indicators. If SC013 is missing, PRIVATESCH is used to create SCHLTYPE.

Similar to 2018, IRL had special treatment for this designation – based solely on the STRATUM sampling variable.

### **Student-teacher ratio (STRATIO)**

The student-teacher ratio (STRATIO) was obtained by dividing the number of enrolled male and female students (SCHSIZE) provided by the principal in SC002 (SC002Q01TA, SC002Q02TA) by the total number of full-time and part-time teachers (TOTAT) provided by the principal in SC018 (SC018Q01TA01, SC018Q01TA02).

### **Student-mathematics teacher ratio (SMRATIO)**

The student-mathematics teacher ratio (SMRATIO) was obtained by dividing the number of enrolled male and female students (SCHSIZE) provided by the principal in SC002 (SC002Q01TA, SC002Q02TA) by the total number of full-time and part-time mathematics teachers (TOTMATH) provided by the principal in SC182 (SC182Q01WA01, SC182Q01WA02).

### **Total number of mathematics teachers at school (TOTMATH)**

Principals were asked in SC182 to report the number of full-time and part-time mathematics teachers at their school (SC182Q01WA01, SC182Q01WA02) and provide additional information on how many of the staff was full-time and part-time employed mathematics teachers qualified at different ISCED levels. The total number of mathematics teachers at the school (TOTMATH) was computed as the sum of full-time and part-time mathematics teachers.

### **Total number of non-teaching staff at school (TOTSTAFF)**

Principals were asked in SC168 to report the number of non-teaching staff currently working in their school. The total number of non-teaching staff at the school (TOTSTAFF) is a sum of the numbers of personnel for pedagogical support (SC168Q01JA), school administrative personnel (SC168Q02JA), school management personnel (SC168Q03JA), and other non-teaching staff (SC168Q04JA).

### **Total number of all teachers at school (TOTAT)**

Principals were asked in SC018 to report the total number of full-time and part-time teachers at their school (SC018Q01TA01, SC018Q01TA02) and provide additional information on how many of the staff was full-time and part-time employed teachers qualified at different ISCED levels.



### *Selection and enrolment (Module 12)*

#### **School selectivity (SCHSEL)**

Principals were asked in SC012 about admittance policies at their school, including student academic performance and recommendation by feeder schools. The three response categories for this question were “Never”, “Sometimes”, and “Always”. An index of academic school selectivity (SCHSEL) was computed by assigning schools to one of three categories based on how often two factors, namely “Student’s record of academic performance” (SC012Q01TA) and “Recommendation of feeder schools” (SC012Q02TA), were considered when admitting students to the school as follows:

1. the two factors (student’s record of academic performance and recommendation of feeder schools) were never considered (if SC012Q01TA=1 and SC012Q02TA=1),
2. at least one of the factors was considered sometimes but neither always (if SC012Q01TA=2 or SC012Q02TA=2, and if SC012Q01TA<3 and SC012Q02TA<3), and
3. at least one of the factors was considered always (if SC012Q01TA=3 or SC012Q02TA=3).

### *School autonomy (Module 13)*

#### **School responsibility for curriculum (SRESPCUR)**

Principals were asked in SC202 about who had the main responsibility for various decisions or activities at their school. The six response categories for this question were “Principal”, “Teachers or members of <school management team>”, “<School governing board>”, “<Local or municipal authority>”, “<Regional or state authority>”, and “<National or federal authority>”. An index of the relative level of responsibility of school staff in deciding issues related to curriculum and assessment (RESPCUR) was computed from the school principals’ reports regarding who had the main responsibility for 4 items in SC202. The index was calculated on the basis of the ratio of responses for “Principal”, “Teachers or members of <school management team>”, or “<School governing board>” on the one hand to responses for “<Local or municipal authority>”, “<Regional or state authority>”, or “<National or federal authority>” on the other hand. In the first step, a count for school responsibility was calculated by counting the number of “Principal”, “Teachers or members of <school management team>”, and “<School governing board>” responses. In the second step, a count for non-school responsibility was calculated by counting the number of “<Local or municipal authority>”, “<Regional or state authority>”, and “<National or federal authority>”. In the third step, the school responsibility count was divided by the non-school responsibility count. To avoid dividing by “0”, “1” was added to both the numerator and denominator; when the ratio of school responsibility to non-school responsibility was 4:0, an index value of 4 was assigned. Higher values indicated relatively higher levels of school responsibility in deciding issues related to curriculum and assessment.

#### **School responsibility for resources (SRESPRES)**

Principals were asked in SC202 about who had the main responsibility for various decisions or activities at their school. The six response categories for this question were “Principal”, “Teachers or members of <school management team>”, “<School governing board>”, “<Local or municipal authority>”, “<Regional or state authority>”, and “<National or federal authority>”. An index of the relative level of responsibility of school staff in deciding issues related to allocating resources (RESPRES) was computed from the school principals’ reports regarding who had the main responsibility for 6 items in SC202. The index was calculated on the basis of the ratio of responses for “Principal”, “Teachers or members of <school management team>”, or “<School governing board>” on the one hand to responses for “<Local or municipal authority>”, “<Regional or state authority>”, or “<National or federal authority>” on the other hand. In the first step, a count for school responsibility was calculated by counting the number of “Principal”,

“Teachers or members of <school management team>”, and “<School governing board>” responses. In the second step, a count for non-school responsibility was calculated by counting the number of “<Local or municipal authority>”, “<Regional or state authority>”, and “<National or federal authority>”. In the third step, the school responsibility count was divided by the non-school responsibility count. To avoid dividing by “0”, “1” was added to both the numerator and denominator; when the ratio of school responsibility to non-school responsibility was 6:0, an index value of 6 was assigned. Higher values on the scale indicated relatively higher levels of school responsibility in this area.

#### *Organisation of student learning at school (Module 14)*

##### **Ability grouping for mathematics classes (ABGMATH)**

School principals were asked in SC187 to report the extent to which their mathematics classes catered to students with different abilities. The three response categories were “For all classes”, “For some classes”, and “Not for any classes”. An index of ability grouping between mathematics classes (ABGMATH) was derived from the first two items (SC187Q01WA, SC187Q02WA) by assigning schools to three categories: (1) schools with no ability grouping for any classes, (2) schools with one of these forms of ability grouping between some classes and (3) schools with one of these forms of ability grouping for all classes.

##### **Class size (CLSIZE)**

Principals were asked in SC003 about the average size of the test language classes in their school. The nine response categories were “15 students or fewer”, “16-20 students”, “21-25 students”, “26-30 students”, “31-35 students”, “36-40 students”, “41-45 students”, “46-50 students”, and “More than 50 students”. The average class size (CLSIZE) was derived from the midpoint of each response category, resulting in a value of 13 for the lowest category, and a value of 53 for the highest.

##### **Math class size (MCLSIZE)**

Principals were asked in SC176 about the average class size of mathematics classes in their school. The nine response categories were “15 students or fewer”, “16-20 students”, “21-25 students”, “26-30 students”, “31-35 students”, “36-40 students”, “41-45 students”, “46-50 students”, and “More than 50 students”. The average math class size (TBD) was derived from the midpoint of each response category, resulting in a value of 13 for the lowest category, and a value of 53 for the highest.

#### *Teacher qualification, training, and professional development (Module 17)*

##### **Proportion of all teachers fully certified (PROATCE)**

Principals were asked in SC018 to report the number of full-time and part-time teachers fully certified by the appropriate authority (SC018Q02TA01, SC018Q02TA02). The proportion of fully certified teachers (PROATCE) was computed by dividing the number of fully certified teachers by the total number of teachers (TOTAT).

##### **Proportion of all teachers with at least ISCED level 6 bachelor qualification (PROPAT6)**

Principals were asked in SC018 to report the number of full-time and part-time teachers with an ISCED level 6 (Bachelor’s or equivalent level) qualification (SC018Q08JA01, SC018Q08JA02). The proportion of teachers with *at least* an ISCED 6 Bachelor qualification (PROPAT6) was calculated by dividing the number full-time and part-time teachers with an ISCED level 6 (Bachelor’s or equivalent level) qualification (SC018Q08JA01, SC018Q08JA02), ISCED level 7 (Master’s or equivalent level) qualification

(SC018Q09JA01, SC018Q09JA02), and ISCED level 8 (Doctoral or equivalent level) qualification (SC018Q10JA01, SC018Q10JA02) by the total number of teachers (TOTAT).

#### **Proportion of all teachers with at least ISCED level 7 master qualification (PROPAT7)**

Principals were asked in SC018 to report the number of full-time and part-time teachers with an ISCED level 7 (Master's or equivalent level) qualification (SC018Q09JA01, SC018Q09JA02). The proportion of teachers with *at least* an ISCED 7 Master qualification (PROPAT7) was calculated by dividing the number of full-time and part-time teachers with an ISCED level 7 (Master's or equivalent level) qualification (SC018Q09JA01, SC018Q09JA02) and ISCED level 8 (Doctoral or equivalent level) qualification (SC018Q10JA01, SC018Q10JA02) by the total number of teachers (TOTAT).

#### **Proportion of all teachers with ISCED level 8 doctoral qualification (PROPAT8)**

Principals were asked in SC018 to report the number of full-time and part-time teachers with an ISCED level 8 (Doctoral or equivalent level) qualification (SC018Q10JA01, SC018Q10JA02). The proportion of teachers with an ISCED 8 Doctoral qualification (PROPAT8) was calculated by dividing the number of these teachers by the total number of teachers (TOTAT).

#### **Proportion of mathematics teachers at school (PROPMATH)**

The proportion of mathematics teachers (PROPMATH) was computed as the total number of full-time and part-time mathematics teachers at their school (TOTMATH) provided by the principal in SC182 (SC182Q01WA01, SC182Q01WA02), divided by the total number of teachers at their school (TOTAT) provided by the principal in SC018 (SC018Q01TA01, SC018Q01TA02).

### *Global crises (Module 21)*

#### **School closure support from education authorities (SCSUPRTD)**

School administrators' responses to three items in SC222 comprise the index on school closure support from education authorities. School administrators who indicated that their school was closed for one or more school days because of COVID-19 were asked to rate the extent that they felt their school was supported by educational authorities during the time that their school building was closed to students because of COVID-19. They reported this information by selecting one of four response options: "Not at all"; "Very little"; "To some extent"; "A lot". Respondents may interpret support broadly to include any kind of assistance (i.e., financial support, volunteer support, etc). If school administrators chose "A lot" to any of the three items, they received a value of "2" on the index. If school administrators chose "Not at all" to all three items, they received a value of "0" on the index. All other responses received a value of "1". The values of the index range from 0-2. This variable was skipped for respondents who reported that their schools had not been closed for COVID-19 on question SC213.

#### **School closure support from other sources (SCSUPRT)**

School administrators' responses to two items in SC222 comprise the index on school closure support from other sources. School administrators who indicated that their school was closed for one or more school days because of COVID-19 were asked to rate the extent that they felt their school was supported by students' parents or guardians and by private donors during the time that their school building was closed to students because of COVID-19. They reported this information by selecting one of four response options: "Not at all"; "Very little"; "To some extent"; "A lot". Respondents may interpret support broadly to include any kind of assistance (i.e., financial support, volunteer support, etc). If school administrators chose "A lot" to either of the two items, they received a value of "2" on the index. If school administrators

chose “Not at all” to both items, they received a value of “0” on the index. All other responses received a value of “1”. The values of the index will range from 0-2. This variable was skipped for respondents who reported that their schools had not been closed for COVID-19 on question SC213.

### *Derived variables based on IRT scaling*

The School Questionnaire provided data for 25 DVs based on IRT scaling. The Cronbach’s alpha for each scale and group are presented in Table 19.127, the number of items with international parameters for each scale and group are presented in Table 19.128, the number of trend items with international parameters for each trend scale and group are presented in Table 19.129, the countries/economies for which the scale scores were suppressed for each scale are presented in Table 19.130, and the groups that did not administer each scale are presented in Table 19.131.

### *School culture and climate (Module 6)*

#### **Negative school climate (NEGSCLIM)**

Principals were asked in SC172 about the extent of problem behaviours that contribute to a negative school climate in their school (e.g., “Profanity”, “Vandalism”). The four response categories for the six items in the scale were “Not at all”, “Small extent”, “Moderate extent”, and “Large extent”. Higher scale score values indicate that problem behaviours contribute to a negative school climate to a greater extent, while lower scale score values indicate that problem behaviours impact school climate to a lesser extent. Table 19.132 shows the item wording and item parameters for the items in this scale.

#### **School diversity and multi-cultural views (DMCVIEWS)**

Principals were asked in SC173 about the school staff’s efforts to promote a diversity-oriented culture and climate during the last academic year (e.g., “They encouraged students of different backgrounds to resolve disagreements by finding common ground.”, “They taught students how to respond to discrimination.”). The five response categories for the six items in the scale were “Never or almost never”, “About once or twice a year”, “About once or twice a month”, “About once or twice a week”, and “Every day or almost every day”. Higher scale score values indicate that diversity-related views were encouraged in the school with greater frequency, while lower scale score values indicate that diversity-related views were encouraged with lesser frequency. Table 19.133 shows the item wording and item parameters for the items in this scale.

#### **Student-related factors affecting school climate (STUBEHA)**

Principals were asked in SC061 about the extent to which student learning is hindered by student behaviours (e.g., “Student truancy”, “Student use of alcohol or illegal drugs”). Note that this scale was linked to the STUBEHA scale in PISA 2018. The four response categories for the six items in the scale were “Not at all”, “Very little”, “To some extent”, and “A lot”. Higher scale score values indicate that students’ learning is hindered to a greater extent by negative student behaviours, while lower scale score values indicate that students’ learning is hindered by negative student behaviours to a lesser extent. Table 19.134 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

#### **Teacher-related factors affecting school climate (TEACHBEHA)**

Principals were asked in SC061 about the extent to which student learning is hindered by teacher behaviours (e.g., “Teacher absenteeism”, “Staff resisting change”). The four response categories for the five items in the scale were “Not at all”, “Very little”, “To some extent”, and “A lot”. Higher scale score values

indicate that students' learning is hindered to a greater extent by negative teacher behaviours, while lower scale score values indicate that students' learning is hindered by negative teacher behaviours to a lesser extent. Table 19.135 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Out-of-school experiences (Module 10)*

#### **Extra-curricular activities offered (ALLACTIV)**

School principals were asked in SC053 to report what extra-curricular activities their schools offered to 15-year-old students (e.g., "School play or school musical", "Mathematics club"). The two response categories for the 10 items in the scale were "Yes" and "No". Higher scale score values indicate that more extra-curricular activities were offered by the school, while lower scale score values indicate that fewer extra-curricular activities were offered. Table 19.136 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *School type and infrastructure (Module 11)*

#### **Shortage of educational material (EDUSHORT)**

Principals were asked in SC017 about the extent to which instruction is hindered by a shortage of educational materials in their school (e.g., "A lack of educational material (e.g., textbooks, IT equipment, library or laboratory material)", "Inadequate or poor quality educational material (e.g. textbooks, IT equipment, library or laboratory material)"). Note that this scale was linked to the EDUSHORT scale in PISA 2018. The four response categories for the four items in the scale were "Not at all", "Very little", "To some extent", and "A lot". Higher scale score values indicate that the school is impacted by a shortage of educational materials to a greater extent, while lower scale score values indicate that the school is impacted to a lesser extent by a shortage of educational materials. Table 19.137 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

#### **Shortage of educational staff (STAFFSHORT)**

Principals were asked in SC017 about the extent to which instruction is hindered by a shortage of educational staff in their school (e.g., "A lack of teaching staff", "Inadequate or poorly qualified assisting staff"). Note that this scale was linked to the STAFFSHORT scale in PISA 2018. The four response categories for the four items in the scale were "Not at all", "Very little", "To some extent", and "A lot". Higher scale score values indicate that the school is impacted by a shortage of educational staff to a greater extent, while lower scale score values indicate that the school is impacted to a lesser extent by a shortage of educational staff. Table 19.138 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *School autonomy (Module 13)*

#### **Educational leadership (EDULEAD)**

Principals were asked in SC201 about how often they or other members of their school management team engaged in activities or behaviours related to educational leadership during the past 12 months (e.g., "Collaborating with teachers to solve classroom discipline problems", "Providing parents or guardians with information on the school and student performance"). The five response categories for the seven items in the scale were "Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", and "Every day or almost every day". Higher scale score values indicate higher frequencies of engagement by the principal and school management team in educational leadership

activities, while lower scale values indicate lower frequencies of engagement by the principal and school management team in educational leadership activities. Table 19.139 shows the item wording and item parameters for the items in this scale.

### **Instructional leadership (INSTLEAD)**

Principals were asked in SC201 about how often they or other members of their school management team engaged in activities or behaviours related to teaching or instructional leadership during the last 12 months (e.g., “Providing feedback to teachers based on observations of instruction in the classroom”, “Taking actions to ensure that teachers feel responsible for their students' learning outcomes”). The five response categories for the five items in the scale were “Never or almost never”, “About once or twice a year”, “About once or twice a month”, “About once or twice a week”, and “Every day or almost every day”. Higher scale score values indicate higher frequencies of engagement by the principal and school management team in instructional leadership activities, while lower scale score values indicate lower frequencies of engagement by the principal and school management team in instructional leadership activities. Table 19.140 shows the item wording and item parameters for the items in this scale.

### **School autonomy (SCHAUTO)**

Principals were asked in SC202 about who had the main responsibility for various decisions or activities at their school (e.g., “Appointing or hiring teachers”, “Determining teachers' salary increases”). The six response categories for the 12 items in the scale were “Principal”, “Teachers or members of <school management team>”, “<School governing board>”, “<Local or municipal authority>”, “<Regional or state authority>”, and “<National or federal authority>”. Higher scale score values indicate that the principal, teachers or members of the school management team, and the school governing board had a greater level of autonomy in decision-making activities at their school. Lower scale score values indicate that these groups had less autonomy. Table 19.141 shows the item wording and item parameters for the items in this scale. It also indicates how the response categories were recoded prior to scaling.

### **Teacher participation (TCHPART)**

Principals were asked in SC202 about who had the main responsibility for various decisions or activities at their school (e.g., “Formulating the school budget”, “Choosing which learning materials are used”). The six response categories for the 12 items in the scale were “Principal”, “Teachers or members of <school management team>”, “<School governing board>”, “<Local or municipal authority>”, “<Regional or state authority>”, and “<National or federal authority>”. Higher scale score values indicate that the teachers or members of the school management team participated to a greater extent in decision-making activities at their school. Lower scale score values indicate that they participated to a lesser extent. Table 19.142 shows the item wording and item parameters for the items in this scale. It also indicates how the response categories were recoded prior to scaling.

## *Organisation of student learning at school (Module 14)*

### **Digital device policies at school (DIGDVPOL)**

School principals were asked in SC190 to indicate whether their school had various policies regarding digital device use (e.g., “Teachers establish rules for when students may use digital devices during lessons.”, “The school has a specific programme to prepare students for responsible internet behaviour.”). The two response categories for the nine items in the scale were “Yes” and “No”. Higher scale score values indicate that digital device policies are enforced at the school to a greater extent, while lower scale score values indicate that such policies are enforced to a lesser extent. Table 19.143 shows the item wording

and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

*Teacher qualification, training, and professional development (Module 17)*

**Mathematics teacher training (MTTRAIN)**

School principals were asked in SC184 to indicate the areas in which professional development was offered to mathematics teachers in their school (e.g., “Mathematics content”, “Mathematics curriculum”). The two response categories for the seven items in the scale were “Yes” and “No”. Higher scale score values indicate that more opportunities for professional development are offered to mathematics teachers in the school, while lower scale score values indicate that fewer professional development opportunities are offered to them. Table 19.144 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

*Assessment, evaluation, and accountability (Module 18)*

**Feedback to teachers (TEAFDBK)**

Principals were asked in SC193 how much impact teacher evaluations had on various matters (e.g., “A change in salary”, “A change in the likelihood of career advancement”). The four response categories for the seven items in the scale were “No impact”, “Small impact”, “Moderate impact”, and “Large impact”. Higher scale score values indicate greater impact of teacher evaluations or feedback, while lower scale score values indicate lesser impact. Table 19.145 shows the item wording and item parameters for the items in this scale.

**Use of standardised tests (STDTEST)**

Principals were asked in SC035 to indicate whether standardised tests were used for various purposes (e.g., “To guide students’ learning”, “To group students for instructional purposes”). The two response categories for the 11 items in the scale were “Yes” and “No”. Higher scale score values indicate that standardised tests are used for accountability purposes to a greater extent, while lower scale score values indicate that these tests are used to a lesser extent. Table 19.146 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

**Use of teacher-developed tests (TDTEST)**

Principals were asked in SC035 to indicate whether teacher-developed tests were used for various purposes (e.g., “To guide students’ learning”, “To group students for instructional purposes”). The two response categories for the 11 items in the scale were “Yes” and “No”. Higher scale score values indicate that teacher-developed tests are used for accountability purposes to a greater extent, while lower scale score values indicate that these tests are used to a lesser extent. Table 19.147 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

*Parental/guardian involvement and support (Module 19)*

**School encouragement of parent or guardian involvement (ENCOURPG)**

Principals were asked in SC192 about how often their school staff engaged parents or guardians in various aspects of students’ educational environment during the last academic year (e.g., “Invited parents or guardians to volunteer for school activities”, “Initiated communications with parents or guardians about

their child’s progress”). The four response categories for the six items in the scale were “Never or almost never”, “A few times a year”, “A few times a month”, and “Once a week or more”. Higher scale score values indicate more frequent efforts by the school staff to engage parents or guardians in becoming involved at the school, while lower scale score values indicate less frequent engagement efforts. Table 19.148 shows the item wording and item parameters for the items in this scale.

### *Creative thinking (Module 20)*

#### **Beliefs about creativity (BCREATSC)**

Principals were asked in SC204 to indicate their level of agreement with statements regarding their beliefs about creativity (e.g., “Creativity can be trained.”, “There are many different ways to be creative.”). The four response categories for the four items in the scale were “Strongly disagree”, “Disagree”, “Agree”, and “Strongly agree”. Higher scale score values indicate that principals endorse, to a greater extent, beliefs about the malleability of creativity and an expansive view of what it means to be creative. Lower scale score values indicate that principals endorse these beliefs to a lesser extent. Table 19.149 shows the item wording and item parameters for the items in this scale.

#### **Creative school activities offered (ACTCRESC)**

Principals were asked in SC207 to indicate how often creative activities are offered in their school (e.g., “Creative writing classes/activities”, “Debate <club>”). The five substantive response categories for the eight items in the scale were “Never or almost never”, “About once or twice a year”, “About once or twice a month”, “About once or twice a week”, and “Every day or almost every day”. There was an additional response category, “Not available at our school”, which was recoded as missing prior to scaling. Higher scale score values indicate a greater frequency of creative activities being offered in school, while lower scale score values indicate creative activities are offered on a less frequent basis. Table 19.150 shows the item wording and item parameters for the items in this scale. It also indicates how the response categories were recoded prior to scaling.

#### **Creative school environment (CREENVSC)**

Principals were asked in SC205 to indicate their level of agreement with statements regarding the encouragement of creative thinking by teachers and through activities at the school (e.g., “Teachers in our school value students’ creativity.”, “Class activities in our school help students think about new ways to solve complex tasks.”). The four response categories for the six items in the scale were “Strongly disagree”, “Disagree”, “Agree”, and “Strongly agree”. Higher scale score values indicate more agreement with the overall view that students’ creativity is encouraged in the school, while lower scale score values indicate less agreement with this view. Table 19.151 shows the item wording and item parameters for the items in this scale.

#### **Openness culture/climate (OPENCUL)**

Principals were asked in SC208 to indicate the extent to which they agree or disagree with statements regarding their students’ orientation towards openness and creativity (e.g., “Most students at my school are creative.”, “Most students at my school enjoy learning new things.”). The four response categories for the nine items in the scale were “Strongly disagree”, “Disagree”, “Agree”, and “Strongly agree”. Higher scale score values indicate that students have a greater orientation towards openness and creativity, while lower scale score values indicate they have less orientation towards openness and creativity. Table 19.152 shows the item wording and item parameters for the items in this scale.



### *Global crises (Module 21)*

Note that the questions in this module were skipped for respondents who reported that their school had not been closed for one or more school days because of COVID-19 in question SC213.

#### **Problems with schools' capacity to provide remote instruction (PROBSCRI)**

School administrators were asked in SC216 to what extent specific challenges hindered their school's capacity to provide remote instruction during the time when the school building was closed to students because of COVID-19 (e.g., "Lack of access to <digital devices> among students", "Lack of learning management systems or school learning platforms (e.g., [Blackboard®], [Edmodo®], [Moodle®], [Google® Classroom™])"). The four response categories for the five items in the scale were "Not at all", "Very little", "To some extent", and "A lot". Higher scale score values indicate a greater level of problems with the schools' capacity to provide remote instruction while the school building was closed to students during the COVID-19 pandemic, while lower scale score values indicate fewer problems with the capacity to provide remote instruction. Table 19.153 shows the item wording and item parameters for the items in this scale.

#### **School preparation for remote instruction – Before pandemic (SCPREBPB)**

School administrators were asked in SC223 whether their school had taken specific actions to prepare for remote instruction (e.g., "Adapting existing curriculum plans for remote instruction (e.g., modifying course requirements, sequence of lessons, grading policies)", "Ensuring that students have access to <digital devices> for remote instruction"). The three response categories for the 10 items in the scale were "Yes, as a standard practice before the COVID-19 pandemic", "Yes, in response to the COVID-19 pandemic", and "No". Prior to scaling, the first response was coded as 1, while the two latter responses were coded as 0. Higher scale score values indicate a higher level of school preparation for remote instruction before the pandemic, while lower scale score values indicate less preparation for remote instruction before the pandemic. Table 19.154 shows the item wording and item parameters for the items in this scale.

#### **School preparation for remote instruction – In response to pandemic (SCPREPAP)**

School administrators were asked in SC223 whether their school had taken specific actions to prepare for remote instruction (e.g., "Adapting existing curriculum plans for remote instruction (e.g., modifying course requirements, sequence of lessons, grading policies)", "Ensuring that students have access to <digital devices> for remote instruction"). The three response categories for the 10 items in the scale were "Yes, as a standard practice before the COVID-19 pandemic", "Yes, in response to the COVID-19 pandemic", and "No". Prior to scaling, the first two responses were coded as 1, while the third response was coded as 0. Higher scale score values indicate a higher level of school preparation for remote instruction after the start of the pandemic, while lower scale score values indicate less preparation for remote instruction after the start of the pandemic. Table 19.155 shows the item wording and item parameters for the items in this scale.

#### **Preparedness for digital learning (DIGPREP)**

School administrators were asked in SC155 to rate their agreement with statements about their school's capacity to use digital devices to enhance learning and teaching ("Teachers have sufficient time to prepare lessons integrating digital devices", "The school has sufficient qualified technical assistant staff"). The four response categories for the six items in the scale were "Strongly disagree", "Disagree", "Agree", and "Strongly agree". Higher scale score values indicate greater capacity for a school to use digital technology for learning and teaching, while lower scale score values indicate less capacity for a school to do so. Table 19.156 shows the item wording and item parameters for the items in this scale.

## Teacher Questionnaire derived variables

The Teacher Questionnaire is an international option that countries/economies could choose to implement. Routing within the questionnaire was used to deliver specific questions to mathematics teachers and specific questions to all other, non-mathematics teachers. Other questions in the questionnaire were administered to all teachers. There were 45 variables derived from this questionnaire, including 13 simple DVs and 32 IRT scaled DVs. An overview of all DVs in this questionnaire is shown in Table 19.157 and each are described in the following sections.

### **Simple questionnaire indices**

#### *Originally trained teachers – Strict (OTT1) and broad (OTT2) definitions*

The Teacher Questionnaire addressed two questions about teachers' initial education and professional development. The first question, TC014, asks if teacher education or training programme was completed, with response options "1, Yes, a programme of 1 year or less" "2, Yes, a programme longer than 1 year" and "3, No". TC015 asked about how the teacher qualification was received. Response options included "1, I attended a standard teacher education or training programme at an <educational institute which is eligible to educate or train teachers>.", "2, I attended an in-service teacher education or training programme.", "3, I attended a work-based teacher education or training programme.", "4, I attended training in another pedagogical profession.", or "5, Other". These two questions (TC014, TC015) were used to build the DV OTT1 (Originally trained teachers, strict definition) and OTT2 (Originally trained teachers, broad 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 an <educational institute which is eligible to educate or train teachers>". In the less strict definition, the teacher has finished any of the following three programmes: either a "standard teacher education or training programme at an <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).

#### *Trained to teach certain subjects (NTEACH1-11)*

TC018 asked 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 DVs NTEACH1 to NTEACH 11 reflect whether the teacher was trained to teach a certain subject. A value of "1" indicate that teachers were trained to teach the subject in question, while "0" indicates they were not trained to teach the subject in question.

#### *Subject-specific overlap between initial education and teaching the modal grade (STTMG1-11)*

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. This question is used to build the DVs 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. A value of "0" indicates that teachers were neither trained nor teach the subject in question, "1" indicates they were trained to teach the subject in question but do not teach it, "2" indicates they were not trained to teach the subject in question but they do teach the subject, and "3" indicates they were trained to teach the subject in question and they teach the subject.

### *Country born (COBN\_T)*

COBN\_T is based on question TC186, which asks about the country/economy a teacher is born in, coded into the following categories: (1) “Country of test” and (2) “Other country”. Each country/economy adapts the items for this question to collect relevant country/economy of birth information for teachers in their country/economy, so the index also gives detailed categories of teachers’ original countries/economies of birth within a country/economy that vary from country/economy to country/economy.

### *Content overlap between initial education and professional development (TC045Q01-TC045Q18)*

TC045 asked about 10 content topics (e.g., “knowledge and understanding of my subject field(s)”, “knowledge of the curriculum”) 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. The DVs TC045Q01 to TC045Q18 reflect the content overlap between initial education and professional development.

### *Higher educational level attained (TCISCED)*

Teachers’ responses to TC210 regarding education were classified using ISCED 2011. An index on higher educational level attained was constructed by recoding educational qualifications into the following categories: Less than ISCED Level 3.3 (upper secondary with no direct access to tertiary education), ISCED Level 3.3 (upper secondary with no direct access to tertiary education), ISCED 3.4 (upper secondary with direct access to tertiary education), ISCED 4 (post-secondary non-tertiary), ISCED 5 (short-cycle tertiary), ISCED 6 (Bachelor’s or equivalent level), ISCED 7 (Master’s or equivalent level), and ISCED 8 (Doctoral or equivalent level). Scores are assigned as noted in Table 19.158.

### *Employment status (EMPLSTAT and EMPLSTATd)*

TC211 asked about employment status in terms of the contract duration with three response options (“Permanent employment”, “Fixed-term contract for a period of more than 1 school year”, “Fixed-term contract for a period of 1 school year or less”). The corresponding DVs reflected the duration of employment, measured via TC211, a) on the original three-point scale (EMPLSTAT) and b) dichotomous, distinguishing a permanent position from fixed-term contracts (EMPLSTATd).

### *Study abroad (STABROAD)*

Teachers’ responses to the whether they had studied abroad in question TC188 were scaled into the index of “Study abroad”. There are four response options (No; Yes, for less than three months; Yes, for three to twelve months; Yes, for more than a year). This question was previously included in the PISA 2018 Teacher Questionnaire, but with no index. Values on this index are 0 (Never studied abroad) and 1 (Studied abroad). This variable was skipped for teachers who reported that they taught mathematics to students in national modal grade for 15-year-olds this school year on question TC217.

### *Weekly teacher workload (TCWKLOAD)*

Teachers were asked to enter how many hours a week they spend on various teaching tasks (e.g., “Marking/correcting of student work”, “General administrative work”). Teachers’ responses to each of the eight fill-in items in question TC216 were summed to create the simple index “Weekly teacher workload”.

### *Use of digital resources for mathematics (ICTMATTC)*

Teachers' frequency ratings of how often they instruct their students to use digital resources for a range of mathematics tasks in class or for homework (e.g., "Use digital resources for simple calculations", "Use digital resources for simulations and modelling, virtual laboratories") in question TC222 were aggregated into a simple index "Use of digital resources for mathematics" as follows: A value of "1" is assigned if teachers select response options 4 ("About once or twice a week") or 5 ("Every day or almost every day") at least once across the 4 items in this question, "0" otherwise. The index captures whether teachers said that they use some digital resources for mathematics lessons at least on a weekly basis or not. This variable was skipped for teachers who reported that they did not teach mathematics to students in national modal grade for 15-year-olds this school year on question TC217.

### *Proportion of working years at this school (PROPWORK)*

In TC007 teachers were asked to how many years of teaching experience they had at the school having them take the questionnaire and how many years of teaching experience they had in total. These responses were used to form the simple index "Proportion of working years at this school", by dividing the number of years at their current school by total number of years teaching.

### ***Derived variables based on IRT Scaling***

The Teacher Questionnaire provided data for 32 DVs based on IRT scaling. The Cronbach's alpha for each scale and group are presented in Table 19.159, the number of items with international parameters for each scale and group are presented in Table 19.160, the number of trend items with international parameters for each trend scale and group are presented in Table 19.161, the countries/economies for which the scale scores were suppressed for each scale are presented in Table 19.162, and the groups that did not administer each scale are presented in Table 19.163.

### *Proportion of professional development (PRPDT)*

Teachers' responses about their participation in different professional development activities in the last 12 months (e.g., "Individual or collaborative research on a topic of interest to you professionally", "Course, workshop, or conference on teaching methods") in question TC020 were scaled into the index of "Proportion of professional development". Each of the 14 items included in this scale had two response options ("Yes", "No"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.164 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Exchange and co-ordination for teaching (EXCHT)*

Teachers' frequency ratings of how often they participate in teaching-related co-operation (e.g., Exchange teaching materials with colleagues", "Attend team conferences") in question TC046 were scaled into the index of "Exchange and co-ordination for teaching". Note that this scale was linked to the EXCHT scale in PISA 2018. Each of the four items included in this scale had six response options ("Never", "Once a year or less", "2-4 times a year", "5-10 times a year", "1-3 times a month", "Once a week or more"). Table 19.165 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Teaching ICT awareness (ICTOTL)*

Teachers' responses about whether they taught various ICT awareness activities to their students (e.g., "How to decide whether to trust information from the Internet", "How to detect phishing or spam emails") in question TC166 were scaled into the index of "Teaching ICT awareness". Each of the seven items included in this scale had two response options ("Yes", "No"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.166 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Teachers' use of specific ICT applications (TCICTUSE)*

Teachers' frequency ratings of how often they used specific ICT applications while teaching (e.g., "Digital learning games", "Data logging and monitoring tools") in question TC169 were scaled into the index of "Teachers' use of specific ICT applications". Note that this scale was linked to the TCICTUSE scale in PISA 2018. Each of the 14 items included in this scale had four response options ("Never", "In some lessons", "In most lessons", "In every or almost every lesson"). Table 19.167 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Disciplinary climate in mathematics (TCDISCLIMA)*

Teachers' frequency ratings of how often a range of situations occurred in their mathematics lessons (e.g., "There is noise and disorder.", "I have to wait a long time for students to quiet down.") in question TC170 were scaled into the index of "Disciplinary climate in mathematics", which measures how much discipline there is during mathematics lessons. Each of the seven items included in this scale had four response options ("Every lesson", "Most lessons", "Some lessons", "Never or almost never"). Note that this scale was skipped for teachers who reported that they did not teach mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.168 shows the item wording and item parameters for the items in this scale.

### *Need for professional development (DEVNEED)*

Teachers' ratings of their need for professional development in various areas (e.g., "Knowledge of the curriculum", "Student behaviour and classroom management") in question TC185 were scaled into the index of "Need for professional development". Each of the 13 items included in this scale had four response options ("No need at present", "Low level of need", "Moderate level of need", "High level of need"). Table 19.169 shows the item wording and item parameters for the items in this scale.

### *Teachers' attitudes toward equal rights for immigrants (TCATTIMM)*

Teachers' ratings of their agreement with statements about immigrants (e.g., "Immigrant children should have the same opportunities for education that other children in the country have.", "Immigrants should have the opportunity to continue their own customs and lifestyle.") in question TC196 were scaled into the index of "Teachers' attitudes toward equal rights for immigrants". This scale was linked to the TCATTIMM scale in PISA 2018. Each of the four items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.170 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Satisfaction with current job environment (SATJOB)*

Teachers' ratings of their agreement with various statements indicating their satisfaction with the current job environment (e.g., "I enjoy working at this school.", "I would recommend my school as a good place to work.") in question TC198 were scaled into the index of "Satisfaction with current job environment". Note that this scale was linked to the SATJOB scale in PISA 2018. Each of the four items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.171 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Satisfaction with teaching profession (SATTEACH)*

Teachers' ratings of their agreement with various statements indicating their satisfaction with the teaching profession (e.g., "The advantages of being a teacher clearly outweigh the disadvantages.", "I regret that I decided to become a teacher.") in question TC198 were scaled into the index of "Satisfaction with teaching profession". Note that this scale was linked to the SATTEACH scale in PISA 2018. Each of the five items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.172 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling and which are trend items.

### *Teacher's self-efficacy in classroom management (SEFFCM)*

Teachers' ratings of their classroom management skills (e.g., "Control disruptive behaviour in the classroom", "Get students to follow classroom rules") in question TC199 were scaled into the index of "Teacher's self-efficacy in classroom management". This scale was linked to the SEFFCM scale in PISA 2018. Each of the four items included in this scale had four response options ("Not at all", "To some extent", "Quite a bit", "A lot"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.173 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Teacher's self-efficacy in maintaining positive relations with students (SEFFREL)*

Teachers' ratings of their ability to maintain positive relations with students (e.g., "Help my students value learning", "Motivate students who show low interest in school work") in question TC199 were scaled into the index of "Teacher's self-efficacy in maintaining positive relations with students". This scale was linked to the SEFFREL scale in PISA 2018. Each of the four items included in this scale had four response options ("Not at all", "To some extent", "Quite a bit", "A lot"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.174 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Teacher's self-efficacy in instructional settings (SEFFINS)*

Teachers' ratings of their self-efficacy in instructional settings (e.g., "Craft good questions for my students", "Provide an alternative explanation for example when students are confused") in question TC199 were scaled into the index of "Teacher's self-efficacy in instructional settings". This scale was linked to the SEFFINS scale in PISA 2018. Each of the four items included in this scale had four response options ("Not at all", "To some extent", "Quite a bit", "A lot"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.175 shows the item wording and item parameters for the items in this scale. It also indicates which items are trend items.

### *Teacher use of ICT (TCDIGRES)*

Teachers' frequency ratings of how often they use ICT for various teaching tasks (e.g., "Use <digital resources> to design tasks", "Use <digital resources> to provide feedback to students") in question TC220 were scaled into the index of "Teacher use of ICT". Each of the nine items included in this scale had five response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day"). Table 19.176 shows the item wording and item parameters for the items in this scale.

### *Emphasis on ICT competencies (ICTCOMP)*

Teachers' ratings of how much emphasis they place on teaching various ICT competencies (e.g., "Evaluating the credibility of digital information", "Using digital tools to work collaboratively") in question TC221 were scaled into the index of "Emphasis on ICT competencies". Each of the five items included in this scale had four response options ("No emphasis", "Little emphasis", "Some emphasis", "A lot of emphasis"). Table 19.177 shows the item wording and item parameters for the items in this scale.

### *Teaching of mathematical reasoning and 21st century mathematics topics (EXPO21TC)*

Teachers' frequency ratings of how often they had taught a range of different mathematics topics during the school year (e.g., "Extracting mathematical information from diagrams, graphs, or simulations", "Using the concept of statistical variation to make a decision") in question TC223 were scaled into the index of "Teaching of mathematical reasoning and 21<sup>st</sup> century mathematics topics". Each of the 10 items included in this scale had four response options ("Frequently", "Sometimes", "Rarely", "Never"). Note that this scale was skipped for teachers who reported that they did not teach mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.178 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Encouraging mathematical thinking (COGACMTC)*

Teachers' frequency ratings of how often they showed a range of behaviours indicative of encouraging mathematical thinking during the school year (e.g., "I encouraged students to "think mathematically".", "I asked students how different topics are connected to a bigger mathematical idea.") in question TC227 were scaled into the index of "Encouraging mathematical thinking". Each of the nine items included in this scale had five response options ("Never or almost never", "Less than half of the lessons", "About half of the lessons", "More than half of the lessons", "Every lesson or almost every lesson"). Note that this scale was skipped for teachers who reported that they did not teach mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.179 shows the item wording and item parameters for the items in this scale.

### *Fostering reasoning (COGACRTC)*

Teachers' frequency ratings of how often they showed a range of behaviours indicative of fostering mathematics reasoning during the school year (e.g., "I asked students to explain their reasoning when solving a mathematics problem.", "I asked students to defend their answer to a mathematics problem.") in question TC228 were scaled into the index of "Fostering reasoning". Each of the nine items included in this scale had five response options. ("Never or almost never", "Less than half of the lessons", "About half of the lessons", "More than half of the lessons", "Every lesson or almost every lesson"). Note that this scale was skipped for teachers who reported that they did not teach mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.180 shows the item wording and item parameters for the items in this scale.

### *Goals and views about teaching mathematics (TCMGOALS)*

Teachers' agreement with statements about their views and goals when teaching mathematics (e.g., "Explaining why an answer is correct is just as important as getting a correct answer.", "Asking students to solve difficult problems in class helps them become good problem solvers.") in question TC230 were scaled into the index of "Goals and views about teaching mathematics". Each of the 11 items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Note that this scale was skipped for teachers who reported that they did not teach mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.181 shows the item wording and item parameters for the items in this scale.

### *Adaptation of instruction (ADAPTINSTR)*

Teachers' frequency ratings of how often they adapt instruction for students (e.g., "I tailor my teaching to meet the needs of my students.", "I provide individual support for advanced students.") in question TC232 were scaled into the index of "Adaptation of instruction". Each of the four items included in this scale had four response options ("Never or almost never", "Some lessons", "Many lessons", "Every lesson or almost every lesson"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.182 shows the item wording and item parameters for the items in this scale.

### *Feedback provided by the teachers (FEEDBINSTR)*

Teachers' frequency ratings of how often they provide feedback to students (e.g., "I give students feedback on their strengths in my course.", "I tell students how they can improve their performance.") in question TC232 were scaled into the index of "Feedback provided by the teachers". Each of the five items included in this scale had four response options ("Never or almost never", "Some lessons", "Many lessons", "Every lesson or almost every lesson"). Note that this scale was skipped for teachers who reported that they taught mathematics to students in the national modal grade for 15-year-olds in the current school year on question TC217. Table 19.183 shows the item wording and item parameters for the items in this scale.

### *Openness to creativity (OPENCTTC)*

Teachers' ratings of their agreement with statements about their openness to creative activities (e.g., "I enjoy projects that require creative solutions.", "I express myself through art.") in question TC234 were scaled into the index of "Openness to creativity". Each of the eight items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.184 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Creative values (CREATVAL)*

Teachers' ratings of their agreement with statements about their values regarding creativity (e.g., "It is important that students are able to make creative works like drawing and painting.", "It is important for students to solve science problems creatively.") in question TC235 were scaled into the index of "Creative values". Each of the six items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.185 shows the item wording and item parameters for the items in this scale.



### *Teachers' use of creative pedagogies (CREATPED)*

Teachers' ratings of how much importance they place on using creative pedagogies in class (e.g., "Finding ideas through brainstorming", "Debating ideas or current issues") in question TC236 were scaled into the index of "Teachers' use of creative pedagogies". Each of the seven items included in this scale had four response options ("No importance", "Very little importance", "Some importance", "A lot of importance"). Table 19.186 shows the item wording and item parameters for the items in this scale.

### *Teachers' capacity to concentrate at work (CAPCON)*

Teachers' frequency ratings of how often they experienced various situations during the school day (e.g., "I was distracted.", "I felt focused.") in question TC237 were scaled into the index of "Teachers' capacity to concentrate at work". Each of the six items included in this scale had four response options ("Never", "Seldom", "Often", "Always"). Table 19.187 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Teachers' affect (AFFECT)*

Teachers' frequency ratings of how often they felt various emotions during the school day (e.g., "I felt cheerful and in good spirits.", "I felt active and vigorous.") in question TC238 were scaled into the index of "Teachers' affect". Each of the five items included in this scale had four response options ("Never", "Seldom", "Often", "Always"). Table 19.188 shows the item wording and item parameters for the items in this scale.

### *Teachers' feeling of trust (TRUST)*

Teachers' ratings of their agreement with statements about a climate of trust within the school (e.g., "Teachers can rely on the school's management for professional support.", "I feel that I can trust my colleagues.") in question TC241 were scaled into the index of "Teachers' feeling of trust". Each of the five items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.189 shows the item wording and item parameters for the items in this scale.

### *Teachers' work overload (OVERLOAD)*

Teachers' ratings of their agreement with statements concerning work overload (e.g., "I am given enough time to do what is expected of me at work.", "I have too much work for one person to do.") in question TC243 were scaled into the index of "Teachers' work overload". Each of the six items included in this scale had four response options ("Strongly disagree", "Disagree", "Agree", "Strongly agree"). Table 19.190 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Teachers' work autonomy (AUTONOMY)*

Teachers' ratings of how much control they have over various decisions at their school (e.g., "Determining course content", "Disciplining students") in question TC246 were scaled into the index of "Teachers' work autonomy". Each of the seven items included in this scale had four response options ("No control", "Some control", "A lot of control", "Full control"). Table 19.191 shows the item wording and item parameters for the items in this scale.

### *School leadership (LEADSHIP)*

Teachers' frequency ratings of how often their school's principal took various actions (e.g., "My principal collaborated with teachers to solve classroom discipline problems.", "My principal observed instruction in the classroom.") in question TC253 were scaled into the index of "School leadership". Each of the seven items included in this scale had four response options ("Never or rarely", "Sometimes", "Often", "Very often"). Table 19.192 shows the item wording and item parameters for the items in this scale.

### *Occupational stress (OCSTRESS)*

Teachers' ratings of their agreement with various statements regarding their stress at work (e.g., "I experience stress in my work.", "My job negatively impacts my mental health.") in question TC254 were scaled into the index of "Occupational stress". Each of the four items included in this scale had four response options ("Not at all", "To some extent", "Quite a bit", "A lot"). Table 19.193 shows the item wording and item parameters for the items in this scale. It also indicates which items were reverse-coded prior to scaling.

### *Sources of stress (STRESS)*

Teachers' ratings of their agreement with various situations causing stress at work (e.g., "Having too many lessons to teach", "Being held responsible for students' achievement") in question TC255 were scaled into the index of "Sources of stress". Each of the nine items included in this scale had four response options ("Not at all", "To some extent", "Quite a bit", "A lot"). Table 19.194 shows the item wording and item parameters for the items in this scale.

### *Negative physical symptoms (NEGSYMPT)*

Teachers' frequency ratings of how often they had various psychosomatic symptoms during the school day (e.g., "Headache", "Fatigue") in question TC256 were scaled into the index of "Negative physical symptoms". Each of the 10 items included in this scale had five response options ("Never or almost never", "About once or twice a year", "About once or twice a month", "About once or twice a week", "Every day or almost every day"). Table 19.195 shows the item wording and item parameters for the items in this scale.

## References

- Avvisati, F. (2020), "The measure of socio-economic status in PISA: A review and some suggested improvements", *Large-Scale Assessments in Education*, Vol. 8/1, pp. 1-37, <https://doi.org/10.1186/s40536-020-00086-x>. [14]
- Bertling, J. and J. Weeks (2020), *Getting More Bang for Your Buck: Within-construct Questionnaire Matrix Sampling*, Paper presented to PISA Technical Advisory Group, September 2020, Princeton, NJ. [3]
- Bertling, J. and J. Weeks (2018), *Plans for Within-construct Questionnaire Matrix Sampling in PISA 2021*, Paper presented to PISA Technical Advisory Group, August 2018, Princeton, NJ. [2]
- Bertling, J. et al. (2020), "Comparison of within-construct matrix sampling with scale shortening: Impact on IRT scaling and population modelling", *Memo prepared for PISA Technical Advisory Group, TAG (2020)2*. [4]

- Birnbaum, A. (1968), "Some latent trait models and their use in inferring an examinee's ability", in Lord, F. and M. Novick (eds.), *Statistical Theories of Mental Test Scores*, Addison-Wesley. [5]
- Cowan, C. et al. (2012), *Improving the measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation*, National Center for Education Statistics, [https://nces.ed.gov/nationsreportcard/pdf/researchcenter/socioeconomic\\_factors.pdf](https://nces.ed.gov/nationsreportcard/pdf/researchcenter/socioeconomic_factors.pdf). [13]
- Ganzeboom, H. and D. Treiman (2003), "Three internationally standardised measures for comparative research on occupational status", in Hoffmeyer-Zlotnik, J. and C. Wolf (eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*, Kluwer Academic Press. [12]
- Masters, G. (1982), "A Rasch model for partial credit scoring", *Psychometrika*, Vol. 47/2, pp. 149-174, <https://doi.org/10.1007/BF02296272>. [10]
- Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", *Applied Psychological Measurement*, Vol. 16/2, pp. 159-177, <https://doi.org/10.1002/j.2333-8504.1992.tb01436.x>. [6]
- OECD (2023), *PISA 2022 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/dfe0bf9c-en>. [1]
- OECD (2020), *PISA 2018 Technical Report*, PISA, OECD Publishing, Paris, <https://www.oecd.org/pisa/data/pisa2018technicalreport/>. [15]
- OECD/Eurostat/UNESCO Institute for Statistics (2015), *ISCED 2011 Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264228368-en>. [16]
- Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Nielsen and Lydiche. [9]
- Shin, H. et al. (2017), *Multidimensional Discrete Latent Trait Models (mdlTM) [Draft manual]*, Educational Testing Service. [7]
- von Davier, M. (2015), *Multidimensional discrete latent trait models (mdlTM) (Version 1.965) [Computer software]*. [8]
- Warm, T. (1989), "Weighted likelihood estimation of ability in item response theory", *Psychometrika*, Vol. 54/3, pp. 427-450, <https://doi.org/10.1007/BF02294627>. [11]

## Notes

1. For the seven trend scales linked to PISA 2012, the Rasch model (Rasch, 1960) was used to scale the dichotomous items, while the partial credit model (PCM) was used to scale the polytomous items, in line with the models used in PISA 2012.
2. The International Standard Classification of Education (ISCED) is used in international educational statistics to classify levels in education systems worldwide. A link to the 2011 framework, ISCED

---

2011, used in PISA 2022 can be found at <https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf>

3. Separate simple DVs (ST250D06JA, ST250D07JA, ST251D08JA, ST251D09JA) combine all the customizations across countries/economies to ST250C06JA, ST250C07JA, ST251C08JA, and ST251C09JA, respectively. See Annex D.
4. A separate simple DV (ST330D10WA) combines all the customizations across countries/economies to ST330C10WA. See Annex D.
5. The random value was drawn from a normal distribution with a mean of 0 and a standard deviation equal to the standard deviation of the residuals of the regression model for the country/economy.
6. Due to missing data within the OECD countries, the senate weights of the individuals that were ultimately included in the calculation of the mean and standard deviation of each component did not sum to 5,000 for each OECD country. Therefore, the OECD countries were only approximately equally weighted in the calculation of the mean and standard deviation of each component.
7. Again, due to missing data within the OECD countries, the senate weights of the individuals that were ultimately included in the calculation of the mean and standard deviation of the preliminary ESCS scores did not sum to 5,000 for each OECD country. Therefore, the OECD countries were only approximately equally weighted in the calculation of the mean and standard deviation of the preliminary ESCS scores.
8. Due to missing data within the OECD countries, the senate weights of the individuals that were ultimately included in the calculation of the mean and standard deviation of each component in PISA 2018 did not sum to 5,000 for each OECD country. Therefore, the OECD countries were only approximately equally weighted in the calculation of the mean and standard deviation of each component in PISA 2018.
9. The random value was drawn from a normal distribution with a mean of 0 and a standard deviation equal to the standard deviation of the residuals of the regression model for the country/economy and cycle.

# Annex 19.A. Methodology and Overview of Derived Variables in PISA 2022 Context Questionnaires

## Annex Table 19.A.1. Chapter 19: Comprehensive Scales and Variables Analysis from PISA 2022 Questionnaires

All tables are available online at the StatLink below the table.

Tables	Title
Table 19.A.2	Student Questionnaire scales using within-construct matrix sampling design
Table 19.A.3	List of all trend scales
Table 19.A.4	OECD mean and standard deviation of the original WLEs for the new scales
Table 19.A.5	OECD mean and standard deviation of the original WLEs in PISA 2018 for trend scales linked to PISA 2018
Table 19.A.6	OECD mean and standard deviation of the original WLEs in PISA 2012 for trend scales linked to PISA 2012
Table 19.A.7	Variables derived from the Student Questionnaire
Table 19.A.8	Mapping of ISCED levels to years of education
Table 19.A.9	ISCEDP values and labels
Table 19.A.10	Creation of EXPECEDU index
Table 19.A.11	Cronbach's alpha for the IRT scales in the Student Questionnaire
Table 19.A.12	Number of items with international parameters for the IRT scales in the Student Questionnaire
Table 19.A.13	Number of trend items with international parameters for the trend IRT scales in the Student Questionnaire
Table 19.A.14	Countries/economies for which the scale scores were suppressed for the IRT scales in the Student Questionnaire
Table 19.A.15	Groups that did not administer the IRT scales in the Student Questionnaire
Table 19.A.16	Items in the HOMEPOS scale
Table 19.A.17	Recoding of items in the HOMEPOS scale
Table 19.A.18	Items in the ICTRES scale
Table 19.A.19	Items in the INFOSEEK scale
Table 19.A.20	Items in the BULLIED scale
Table 19.A.21	Items in the FEELSAFE scale
Table 19.A.22	Items in the TEACHSUP scale
Table 19.A.23	Items in the RELATST scale
Table 19.A.24	Items in the SCHRISK scale
Table 19.A.25	Items in the BELONG scale
Table 19.A.26	Items in the GROSAGR scale
Table 19.A.27	Items in the ANXMAT scale
Table 19.A.28	Items in the MATHEFF scale
Table 19.A.29	Items in the MATHEF21 scale
Table 19.A.30	Items in the MATHPERS scale
Table 19.A.31	Items in the FAMCON scale
Table 19.A.32	Items in the ASSERAGR scale
Table 19.A.33	Percent of students that did not receive a scale score for ASSERAGR due to extreme straightlining or for not having enough responses
Table 19.A.34	Items in the COOPAGR scale
Table 19.A.35	Percent of students that did not receive a scale score for COOPAGR due to extreme straightlining or for not having enough responses
Table 19.A.36	Items in the CURIOAGR scale
Table 19.A.37	Percent of students that did not receive a scale score for CURIOAGR due to extreme straightlining or for not having enough responses

Tables	Title
Table 19.A.38	Items in the EMOCOAGR scale
Table 19.A.39	Percent of students that did not receive a scale score for EMOCOAGR due to extreme straightlining or for not having enough responses
Table 19.A.40	Items in the EMPATAGR scale
Table 19.A.41	Percent of students that did not receive a scale score for EMPATAGR due to extreme straightlining or for not having enough responses
Table 19.A.42	Items in the PERSEVAGR scale
Table 19.A.43	Percent of students that did not receive a scale score for PERSEVAGR due to extreme straightlining or for not having enough responses
Table 19.A.44	Items in the STRESAGR scale
Table 19.A.45	Percent of students that did not receive a scale score for STRESAGR due to extreme straightlining or for not having enough responses
Table 19.A.46	Items in the EXPOFA scale
Table 19.A.47	Items in the EXPO21ST scale
Table 19.A.48	Items in the COGACRCO scale
Table 19.A.49	Items in the COGACMCO scale
Table 19.A.50	Items in the DISCLIM scale
Table 19.A.51	Items in the FAMSUP scale
Table 19.A.52	Items in the CREATFAM scale
Table 19.A.53	Items in the CREATSCH scale
Table 19.A.54	Items in the CREATEFF scale
Table 19.A.55	Items in the CREATOP scale
Table 19.A.56	Items in the IMAGINE scale
Table 19.A.57	Items in the OPENART scale
Table 19.A.58	Items in the CREATAS scale
Table 19.A.59	Items in the CREATOOS scale
Table 19.A.60	Items in the FAMSUPSL scale
Table 19.A.61	Items in the FEELLAH scale
Table 19.A.62	Items in the PROBELF scale
Table 19.A.63	Items in the SDLEFF scale
Table 19.A.64	Items in the SCHSUST scale
Table 19.A.65	Items in the LEARRES scale
Table 19.A.66	OECD mean and standard deviation of each component of ESCS and the preliminary ESCS scores for PISA 2022
Table 19.A.67	Variables derived from the Financial Literacy Questionnaire
Table 19.A.68	Cronbach's alpha for the IRT scales in the Financial Literacy Questionnaire
Table 19.A.69	Number of items with international parameters for the IRT scales in the Financial Literacy Questionnaire
Table 19.A.70	Number of trend items with international parameters for the trend IRT scales in the Financial Literacy Questionnaire
Table 19.A.71	Groups that did not administer the IRT scales in the Financial Literacy Questionnaire
Table 19.A.72	Items in the FLSCHOOL scale
Table 19.A.73	Items in the FLMULTSB scale
Table 19.A.74	Items in the FLFAMILY scale
Table 19.A.75	Items in the ACCESSFP scale
Table 19.A.76	Items in the FLCONFIN scale
Table 19.A.77	Items in the FLCONICT scale
Table 19.A.78	Items in the ACCESSFA scale
Table 19.A.79	Items in the ATTCONFM scale
Table 19.A.80	Items in the FRINFLFM scale
Table 19.A.81	Variables derived from the ICT Familiarity Questionnaire
Table 19.A.82	Cronbach's alpha for the IRT scales in the ICT Familiarity Questionnaire
Table 19.A.83	Number of items with international parameters for the IRT scales in the ICT Familiarity Questionnaire
Table 19.A.84	Countries/economies for which the scale scores were suppressed for the IRT scales in the ICT Familiarity Questionnaire
Table 19.A.85	Groups that did not administer the IRT scales in the ICT Familiarity Questionnaire
Table 19.A.86	Items in the ICTSCH scale
Table 19.A.87	Items in the ICTHOME scale

Tables	Title
Table 19.A.88	Items in the ICTQUAL scale
Table 19.A.89	Items in the ICTSUBJ scale
Table 19.A.90	Items in the ICTENQ scale
Table 19.A.91	Items in the ICTFEED scale
Table 19.A.92	Items in the ICTOUT scale
Table 19.A.93	Items in the ICTWKDY scale
Table 19.A.94	Items in the ICTWKEND scale
Table 19.A.95	Items in the ICTREG scale
Table 19.A.96	Items in the ICTINFO scale
Table 19.A.97	Items in the ICTEFFIC scale
Table 19.A.98	Variables derived from the Well-Being Questionnaire
Table 19.A.99	Cronbach's alpha for the IRT scales in the Well-Being Questionnaire
Table 19.A.100	Number of items with international parameters for the IRT scales in the Well-Being Questionnaire
Table 19.A.101	Number of trend items with international parameters for the trend IRT scales in the Well-Being Questionnaire
Table 19.A.102	Groups that did not administer the IRT scales in the Well-Being Questionnaire
Table 19.A.103	Items in the BODYIMA scale
Table 19.A.104	Items in the SOCONPA scale
Table 19.A.105	Items in the LIFESAT scale
Table 19.A.106	Items in the PSYCHSYM scale
Table 19.A.107	Items in the SOCCON scale
Table 19.A.108	Items in the EXPWB scale
Table 19.A.109	Variables derived from the Parent Questionnaire
Table 19.A.110	Creation of the PAREXPT index
Table 19.A.111	Cronbach's alpha for the IRT scales in the Parent Questionnaire
Table 19.A.112	Number of items with international parameters for the IRT scales in the Parent Questionnaire
Table 19.A.113	Number of trend items with international parameters for the trend IRT scales in the Parent Questionnaire
Table 19.A.114	Countries/economies for which the scale scores were suppressed for the IRT scales in the Parent Questionnaire
Table 19.A.115	Groups that did not administer the IRT scales in the Parent Questionnaire
Table 19.A.116	Items in the CURSUPP scale
Table 19.A.117	Items in the PQMIMP scale
Table 19.A.118	Items in the PQMCAR scale
Table 19.A.119	Items in the PARINVOL scale
Table 19.A.120	Items in the PQSCHOOL scale
Table 19.A.121	Items in the PASCHPOL scale
Table 19.A.122	Items in the ATTIMMP scale
Table 19.A.123	Items in the CREATHME scale
Table 19.A.124	Items in the CREATACT scale
Table 19.A.125	Items in the CREATOPN scale
Table 19.A.126	Items in the CREATOR scale
Table 19.A.127	Variables derived from the School Questionnaire
Table 19.A.128	Cronbach's alpha for the IRT scales in the School Questionnaire
Table 19.A.129	Number of items with international parameters for the IRT scales in the School Questionnaire
Table 19.A.130	Number of trend items with international parameters for the trend IRT scales in the School Questionnaire
Table 19.A.131	Countries/economies for which the scale scores were suppressed for the IRT scales in the School Questionnaire
Table 19.A.132	Groups that did not administer the IRT scales in the School Questionnaire
Table 19.A.133	Items in the NEGSCCLIM scale
Table 19.A.134	Items in the DMCVIEWS scale
Table 19.A.135	Items in the STUBEHA scale
Table 19.A.136	Items in the TEACHBEHA scale
Table 19.A.137	Items in the ALLACTIV scale
Table 19.A.138	Items in the EDUSHORT scale
Table 19.A.139	Items in the STAFFSHORT scale
Table 19.A.140	Items in the EDULEAD scale

Tables	Title
Table 19.A.141	Items in the INSTLEAD scale
Table 19.A.142	Items in the SCHAUTO scale
Table 19.A.143	Items in the TCHPART scale
Table 19.A.144	Items in the DIGDVPOL scale
Table 19.A.145	Items in the MTTRAIN scale
Table 19.A.146	Items in the TEAFDBK scale
Table 19.A.147	Items in the STDTEST scale
Table 19.A.148	Items in the TDTEST scale
Table 19.A.149	Items in the ENCOURPG scale
Table 19.A.150	Items in the BCREATSC scale
Table 19.A.151	Items in the ACTCRESC scale
Table 19.A.152	Items in the CREENVSC scale
Table 19.A.153	Items in the OPENCUL scale
Table 19.A.154	Items in the PROBSCRI scale
Table 19.A.155	Items in the SCPREPBP scale
Table 19.A.156	Items in the SCPREPAP scale
Table 19.A.157	Items in the DIGPREP scale
Table 19.A.158	Variables derived from the Teacher Questionnaire
Table 19.A.159	Creation of the TCISCED index
Table 19.A.160	Cronbach's alpha for the IRT scales in the Teacher Questionnaire
Table 19.A.161	Number of items with international parameters for the IRT scales in the Teacher Questionnaire
Table 19.A.162	Number of trend items with international parameters for the trend IRT scales in the Teacher Questionnaire
Table 19.A.163	Countries/economies for which the scale scores were suppressed for the IRT scales in the Teacher Questionnaire
Table 19.A.164	Groups that did not administer the IRT scales in the Teacher Questionnaire
Table 19.A.165	Items in the PRPDT scale
Table 19.A.166	Items in the EXCHT scale
Table 19.A.167	Items in the ICTOTL scale
Table 19.A.168	Items in the TCICTUSE scale
Table 19.A.169	Items in the TCDISCLIMA scale
Table 19.A.170	Items in the DEVNEED scale
Table 19.A.171	Items in the TCATTIMM scale
Table 19.A.172	Items in the SATJOB scale
Table 19.A.173	Items in the SATTEACH scale
Table 19.A.174	Items in the SEFFCM scale
Table 19.A.175	Items in the SEFFREL scale
Table 19.A.176	Items in the SEFFINS scale
Table 19.A.177	Items in the TCDIGRES scale
Table 19.A.178	Items in the ICTCOMP scale
Table 19.A.179	Items in the EXPO21TC scale
Table 19.A.180	Items in the COGACMTC scale
Table 19.A.181	Items in the COGACRTC scale
Table 19.A.182	Items in the TCMGOALS scale
Table 19.A.183	Items in the ADAPTINSTR scale
Table 19.A.184	Items in the FEEDBINSTR scale
Table 19.A.185	Items in the OPENCTTC scale
Table 19.A.186	Items in the CREATVAL scale
Table 19.A.187	Items in the CREATPED scale
Table 19.A.188	Items in the CAPCON scale
Table 19.A.189	Items in the AFFECT scale
Table 19.A.190	Items in the TRUST scale
Table 19.A.191	Items in the OVERLOAD scale
Table 19.A.192	Items in the AUTONOMY scale
Table 19.A.193	Items in the LEADSHIP scale



Tables	Title
Table 19.A.194	Items in the OCSTRESS scale
Table 19.A.195	Items in the STRESS scale
Table 19.A.196	Items in the NEGSYMPT scale

StatLink  <https://stat.link/v6uq1n>

# **20** Questionnaire Design and the Computer-Based Questionnaire Platform

## Introduction

Questionnaires are a critical component of the PISA survey, providing important information about the context in which students learn and live as well as demographics and other reporting information. When coupled with the cognitive results, the questionnaires can provide insights into relationships between background information and cognitive achievement. The mode of administration of the questionnaires has evolved across the PISA cycles. PISA administered questionnaires in paper-based format alone until PISA 2012 when an optional online School Questionnaire was introduced. Computer-based questionnaires were implemented more broadly in PISA 2015 and expanded in PISA 2018 to be the primary mode of questionnaire administration. In PISA 2022, computer-based questionnaires continued to be the primary mode of administration and all questionnaires except the Parent Questionnaire were available in computer-based format. The use of a computer-based questionnaire administration allowed for innovations to be introduced in the PISA 2022 cycle, as well as continuing to increase the data quality over PISA 2018.

Annex Table 20.A.2. shows the compulsory and optional questionnaires that were administered in PISA 2022 and their administration mode.

This chapter first explains the PISA 2022 questionnaire design for the Field Trial and the Main Survey, then provides an overview of the process used to author the international master and the national questionnaires, and finally provides an overview of the technical design of the questionnaire platform.

## Questionnaire Design

PISA emphasizes the importance of collecting context information from students and schools along with the assessment of student achievement. A Student Questionnaire (STQ) and a School Questionnaire (SCQ) cover a broad range of contextual variables. The content of these questionnaires – especially the content of the Student Questionnaire – changes considerably between cycles based on the major domain of the assessment, but the administration has remained stable: every student participating in the PISA assessment completes the STQ, and every school principal of the participating schools, one per school, completes the SCQ.

PISA has also included several international questionnaire options, i.e. additional instruments that countries could administer as an international option. For PISA 2022, it included a Parent Questionnaire (PAQ), Teacher Questionnaire (TQ), and optional questionnaires for the students: the Financial Literacy Questionnaire (FLQ), ICT Familiarity Questionnaire (ICQ), and Well-Being Questionnaire (WBQ). Annex

Table 20.A.5. (from Chapter 1) summarises the participation of countries/economies in the different international questionnaires.

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 parameters for each participating country. In addition, the design also must ensure that important policy issues and research questions can be addressed in later analysis and reporting based on PISA 2022 data. Both the psychometric quality of the variables and indicators and the analytical power of the study must be considered when proposing and evaluating a questionnaire design. This is usually done by pre-testing all questionnaire content and innovations 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 in the Main Survey. Results are then discussed with the PISA expert groups and Main Survey material is selected.

The Field Trial and the Main Study questionnaire designs differ greatly in many respects. The goal of the PISA 2022 Field Trial questionnaires is to re-evaluate the quality of the context questionnaire items used in previous cycles as well as the quality of new items developed for this cycle and test methodological innovations, namely the within-construct matrix sampling of items. Moreover, the PISA 2022 Field Trial provided an opportunity for countries to test the questionnaire administration procedures. The main survey questionnaires must collect equivalent information across all students, schools, teachers, and parents to be used in reporting results, and thus the design needs to administer the same questions to all respondents.

The following sections discuss the main differences between the PISA 2022 Field Trial and the Main Survey design for both the paper-based and computer-based questionnaires.

## Student-Administered Questionnaires

In the field trial, approximately 1 992 students per country respond to the computer-based questionnaires and approximately 900 students per country respond to the paper-based questionnaires. These numbers were increased in the main survey to approximately 6 300 students per computer-based country and 5 250 students per paper-based country. In addition, countries/economies opting to administer the Financial Literacy assessment sampled an additional 1 650 students. Because of the differences in the tools available for authoring a computer-based questionnaire versus a paper-based questionnaire, the design of the computer-based and paper-based questionnaires administered to students were different.

### **Computer-based design**

#### *Field trial*

The field trial Student Questionnaire design for PISA 2022 allowed for the maximum number of potential items to be tested, experiments on different wording and formats of similar questions, and the piloting of the new within-construct matrix sampling methodology. To accomplish these goals, the Student Questionnaire design included two overlapping virtual booklets, each of which included a subset of items with different wording for version comparison experiments that were administered to only half the students taking that booklet. This resulted in four major paths through the questionnaire: Booklets 1a, 1b, 2a, and 2b. The field trial Student Questionnaire design is shown in Table 20.1. and describes the constructs included in each of the booklets and experiments. The field trial Student Questionnaire was authored as one form and each student had an equal chance of being assigned to one of the four paths.

Table 20.1. Field trial computer-based design for Student Questionnaire

Student Questionnaire			
<b>Common items:</b> basic demographics (age, gender), educational career, migration and language exposure, ESCS (home possessions, guardians)			
<b>Booklet 1:</b> ESCS: Parental education and occupation (parent/guardian versions)		<b>Booklet 2:</b> ESCS: Parental education and occupation (trend mother/father versions)	
<b>Common items:</b> ESCS (food insecurity and subjective socioeconomic status)			
<b>Booklet 1:</b> <ul style="list-style-type: none"> <li>• Educational career</li> <li>• School culture and climate</li> <li>• Out-of-school experiences</li> <li>• Parent/guardian involvement and support</li> <li>• Social and emotional characteristics (vignettes)</li> <li>• Health and well-being</li> <li>• Postsecondary preparedness and aspirations</li> <li>• Future aspirations</li> </ul>		<b>Booklet 2:</b> <ul style="list-style-type: none"> <li>• Migration and language exposure</li> <li>• Organisation of student learning at school</li> <li>• Mathematics teacher behaviours</li> <li>• Exposure to mathematics content</li> <li>• School culture and climate</li> <li>• Subject-specific beliefs, attitudes, feelings, and behaviours</li> <li>• Out-of-school experiences</li> <li>• Organisation of student learning at school</li> <li>• Creative Thinking</li> </ul>	
<b>Booklet 1a:</b> Social and Emotional Characteristics (agreement)	<b>Booklet 1b:</b> Social and Emotional Characteristics (frequency)	<b>Booklet 2a:</b> <ul style="list-style-type: none"> <li>• Mathematics teacher behaviours (version A)</li> <li>• Exposure to mathematics content (version A)</li> <li>• Subject-specific beliefs, attitudes, feelings, and behaviours (version A)</li> </ul>	<b>Booklet 2b:</b> <ul style="list-style-type: none"> <li>• Mathematics teacher behaviours (version B)</li> <li>• Exposure to mathematics content (version B)</li> <li>• Subject-specific beliefs, attitudes, feelings, and behaviours (version B)</li> </ul>
<b>Common items:</b> Future aspirations, Global crises, PISA preparation and effort			
Financial Literacy Questionnaire			
ICT Familiarity Questionnaire			
WBQ			

The PISA 2022 field trial Student Questionnaire also tested the implementation of the within-construct matrix sampling for the first time in PISA. Certain PISA constructs are measured using 6 to 12 individual items. However, to decrease the amount necessary for responding to each instrument, each student is administered only 5 of the items from a construct. To this effect, each student received a random selection of the items in a construct, so students received different combinations of the items from the construct. The computer platform used a combination of the student questionnaire random number and the position of the question screen within the questionnaire form to determine which items to show so that students were shown items in different positions on each screen (e.g. a student did not always see items 1, 2, 4, 5, and 7 on every matrix-sampled screen). The master version of the field trial questionnaire identified all the questions for which the within-construct matrix sampling would be applied.

The optional questionnaires for students: Financial Literacy (FLQ), ICT Familiarity (ICQ), and Well-being (WBQ) were administered following the Student Questionnaire and were available only as computer-based instruments. These optional questionnaires each consisted of a single form with no virtual booklets or version comparisons. The field trial version of the FLQ and ICQ contained more content than needed for the main study in order to evaluate the quality of new items. The WBQ was administered unchanged from PISA 2018. Within-construct matrix sampling was not applied to any questions in the optional questionnaires.

The computer-based Student Questionnaire Une Heure (STQ-UH) booklet consisted of a subset of questions from the Student Questionnaire designed to take approximately 20 minutes for a student to

complete. If students received the STQ-UH, they did not receive any of the other optional questionnaires even if their country had elected to participate in those options.

### *Main survey*

For the main survey, the number of items in the Student Questionnaire was reduced significantly as decisions were made about which version of each of the experimental wording questions collected the highest quality data and which of the other new questions collected the highest quality data on other constructs. The main survey Student Questionnaire consisted of one booklet of items that was administered to all students. In creating the main survey questionnaire, the sequence of items was updated because items from two virtual booklets were combined into one virtual booklet. Within-construct matrix sampling was still applied to those questions where the field trial analysis showed that high-quality scales could be constructed from the matrix-sampled items.

The design for the optional Financial Literacy, ICT Familiarity, Well-being, and Parent Questionnaires remained the same as in the Field Trial with one form per questionnaire and the number of items administered in the FLQ and ICQ was reduced slightly to eliminate items that were deemed to not function well, and to reduce the time needed to complete each questionnaire.

## **Paper-based design**

### *Field trial*

A paper-based Student Questionnaire was administered in the four countries that chose the paper-based mode of delivery for both the questionnaires and the cognitive assessment. The paper-based Student Questionnaire took up to 41 minutes of assessment time and included a subset of the items from the field trial computer-based version. The paper-based version did not include questions on the creative thinking module, eliminated some of the version comparison experiments, and had a reduced item pool for a few modules. The field trial paper-based Student Questionnaire was administered in two overlapping booklets as shown in Table 20.2. Students were randomly assigned to one of the two booklets during the survey administration. Within-construct matrix sampling could not be applied to the paper-based version, so students taking the questionnaire on paper received all items in a question instead of a random subset. This did not increase the response time since the paper-based version had fewer items than the computer-based version.

Where possible, questions were administered in the same format and layout in both the paper-based and computer-based questionnaires. However, some questions had to be changed from drop-down or slider response format to open-ended format to accommodate data collection in the PBA mode.

**Table 20.2. Field Trial Paper-based Design for Students**

<b>Paper-based Student Questionnaire</b>	
<b>Common items:</b> <ul style="list-style-type: none"> <li>• Basic demographics (grade, age, gender)</li> <li>• Educational career</li> <li>• Migration and language exposure</li> <li>• Organisation of student learning at school</li> <li>• ESCS: home possessions</li> </ul>	
<b>Booklet 1:</b> <ul style="list-style-type: none"> <li>• ESCS: Parental education and occupation (new parent/guardian versions)</li> <li>• Educational career</li> <li>• Mathematics teacher behaviours (version A)</li> </ul>	<b>Booklet 2:</b> <ul style="list-style-type: none"> <li>• ESCS: Parental education and occupation (trend mother/father versions)</li> <li>• ESCS: Food insecurity and subjective socioeconomic status</li> <li>• Migration and language exposure</li> </ul>

<ul style="list-style-type: none"> <li>• School culture and climate</li> <li>• Exposure to mathematics content</li> <li>• Subject-specific beliefs, attitudes, feelings, and behaviours</li> <li>• Participation in additional mathematics instruction</li> <li>• Social and emotional characteristics (agreement)</li> <li>• Social and emotional characteristics (vignettes)</li> </ul>	<ul style="list-style-type: none"> <li>• Educational Career</li> <li>• School Culture and Climate</li> <li>• Mathematics teacher behaviours (version B)</li> <li>• Out-of-school experiences</li> <li>• Parental involvement and support</li> <li>• Social and emotional characteristics (frequency)</li> <li>• Health and well-being</li> <li>• Post-secondary preparedness and aspirations</li> </ul>
<b>Common items:</b> Global crises, PISA preparation and effort	

International option questionnaires were not available to paper-based countries, so the FLQ, ICQ, and WBQ were not administered to these participants. In addition, no paper-based country chose to administer the STQ-UH questionnaire.

### *Main Survey*

For the main survey, items that were removed from the field trial version of the computer-based student questionnaire were also removed from the paper-based version. The paper-based student questionnaire content was combined into one 35-minute booklet that was administered to all students. When combining the questions from multiple booklets, the sequence of questions was updated as well.

## School Questionnaire

Each school that participated in PISA completed one School Questionnaire to provide contextual information on the environment in which students learn. Since the school questionnaire was answered by approximately 28 schools in the Field Trial, there was not a large enough sample size to administer two booklets of material in the field trial and still evaluate the quality of the items in each country. Therefore, the design of the school questionnaire remained consistent between the field trial and the main survey for both computer-based and paper-based administration, with slightly more material administered in the field trial than in the main survey.

### **Field Trial**

The School Questionnaire in the Field Trial included trend and new material and took approximately 60 minutes to complete. This questionnaire was designed as one form without virtual booklets or version comparisons, and the same questions were administered in both the paper-based and computer-based versions. Data quality was improved in the computer-based version by using automated checks and routing. Certain questions in the computer-based version contained automated range limits that would not permit unrealistic values to be entered and soft checks to encourage the respondent to confirm responses that were higher or lower than expected. In the computer-based version, when filter questions were implemented, routing rules could be used to automatically hide skipped questions from respondents. No automatic range checks are possible in the paper-based version, but rather these were implemented during data entry. Also, in the paper-based version, respondents saw printed instructions to skip certain questions based on their response to a filter question. The computer-based questionnaire also contained four questions where data was collected using sliders instead of the open-response format used in the paper-based questionnaire.

### **Main Survey**

For the Main Survey, the number of questions administered in the School Questionnaire was reduced to 45 minutes of material. The same questions were administered to both the computer-based and paper-

based administration countries. The same technical differences between the computer and paper-based administration were present for the main survey. As the main survey collects information from a larger sample than in the field trial, there are approximately 150 respondents per country to the School Questionnaire.

## Teacher Questionnaire

The Teacher Questionnaire is administered to up to 25 respondents per school. The sampling process attempts to have an equal number of teachers of the major PISA domain (mathematics) and teachers of other subject areas. Additional information about the sampling of teachers can be found in Chapter 6. Certain items in the questionnaire are domain-specific and designed to collect information only from teachers of mathematics, and other questions are general and may be answered by all teachers of 15-year-olds. In the PISA 2018 cycle, the Teacher Questionnaire was designed as two independent forms (one for teachers of reading as a subject, and one for teachers of all other subject areas), and teachers were instructed to log into the version of the questionnaire that applied to them. Once a teacher logged into one form of the questionnaire, they were not able to switch to the other form, and so if a teacher was erroneously classified as a major-domain teacher then they would be asked to respond to questions that did not apply to them. This design was changed for the PISA 2022 cycle, and in this cycle the Teacher Questionnaire was administered as one form and teachers were asked to self-identify as teachers of mathematics or teachers of other subjects and questions were presented based on those responses. If a teacher erroneously marked that they were a teacher of mathematics, they could go back and change their answer and then route to the appropriate questions for their subject area.

### ***Field Trial***

The optional computer-based Teacher Questionnaire was designed as one form containing two overlapping virtual booklets of questions and took approximately 60 minutes to complete in the Field Trial. All teachers first answered a set of common questions about their background, including whether they are a maths teacher or not. Then teachers were routed to one of two blocks of questions: maths-specific questions for teachers who self-identified as teachers of mathematics or general questions for teachers who self-identified as not teaching mathematics. Table 20.3. shows the Teacher Questionnaire Field Trial design.

**Table 20.3. Field trial computer-based design for Teacher Questionnaires (TCQ)**

<b>Teacher Questionnaire</b>	
<b>Common items:</b> <ul style="list-style-type: none"> <li>• Socio-demographic characteristics</li> <li>• Education, Certification, Teacher qualifications</li> <li>• Employment status, years of experience, grade assigned</li> <li>• Workload</li> <li>• Alignment of roles, content of teacher preparation</li> <li>• Support for learning and development</li> <li>• Self-reported impact of professional development on teaching practices</li> <li>• Use of specific ICT applications</li> <li>• Assessments, evaluation and feedback</li> <li>• Emphasis on ICT competencies</li> <li>• Job satisfaction</li> </ul>	
<b>Administered to self-identified mathematics teachers:</b> <ul style="list-style-type: none"> <li>• Teacher qualifications</li> <li>• Mathematics curriculum</li> <li>• Disciplinary climate in mathematics</li> <li>• Use of digital devices in mathematics</li> <li>• Exposure to formal and applied mathematics tasks</li> <li>• Exposure to mathematics problems requiring reasoning</li> <li>• Structure of mathematics instruction</li> <li>• Cognitive activation in mathematics</li> <li>• Mathematics teacher feedback</li> <li>• Goals and views about teaching mathematics</li> <li>• Support for learning and development</li> </ul>	<b>Administered to self-identified non-mathematics teachers:</b> <ul style="list-style-type: none"> <li>• Teacher qualifications</li> <li>• Self-efficacy</li> <li>• Students' engagement with their learning</li> <li>• Opportunity to learn</li> <li>• Attitudes toward immigrants</li> <li>• Teacher background – studying experience abroad</li> </ul>
<b>Common items:</b> <ul style="list-style-type: none"> <li>• Creative thinking</li> <li>• Teacher well-being</li> <li>• Classroom composition</li> </ul>	

### **Main Survey**

The main survey Teacher Questionnaire design was unchanged from the field trial: it consisted of common questions administered to all teachers as well as distinct blocks of questions that were administered to either mathematics teachers or teachers of other subjects. The questionnaires in total still covered all policy modules proposed in the questionnaire framework for this cycle (see Chapter 5). However, the number of questions administered was reduced to eliminate those questions that had lower data quality and/or were not as critical for the purpose of PISA so that the total main survey questionnaire response time was approximately 45 minutes.

### **Parent Questionnaire**

The optional Parent Questionnaire (PAQ) was administered on paper only. Only one Parent Questionnaire was completed for each student, and the questionnaire could be completed by either parent or the student's guardian. The PAQ included trend items as well as newly developed content. The field trial version of the Parent Questionnaire contained approximately 35 minutes of material so that new content could be considered for inclusion in PISA 2022. The main survey version of the Parent Questionnaire contained approximately 30 minutes of material.



## Computer-based Questionnaire Platform

The computer-based questionnaires were designed and administered using the PISA Questionnaire Authoring Tool (QAT), a platform focused on the specific goal of production (i.e. the definition, authoring, testing, translation, adaptation, and validation) of the Master and National versions of the questionnaires, the delivery of these questionnaires to the appropriate respondents, and the management of all administrative tasks relating to questionnaire delivery.

The QAT editor is used to create the questions, routing logic, and consistency checks used in the computer-based questionnaires of PISA 2022. It is an online editor that allows administrators to create a profile for each questionnaire for each country, and then allows users to add, delete, or edit a questions and routings within those questionnaires. Users edit the content and question format of items in the questionnaire in the Editor, then that structure is transformed by the platform into the formatted screens presented to users (the runtime version), and finally the translation of the text is integrated into the runtime to show the questions in each national language.

When users log into the QAT, they are taken to the home page shown in Figure 20.1. This page gives users access to the many features of the tool.

Figure 20.1. Questionnaire Authoring Tool home page

**PISA 2022 Questionnaires** [Log out](#)

**HOME** MONITORING AND DATA ACCESS ADMINISTRATION

**View, edit and test questionnaires**

Here you can preview, test and edit the questionnaires.

Please choose a questionnaire to view or edit:

- [Questionnaire Authoring Tool](#)
- [Review and Test the Questionnaire](#)
- [Upload XLIFF for Preview](#)
- [Clear Testing Results](#)
- [Export Testing Results](#)
- [Export Spreadsheet](#)

**Copy Items Between Questionnaires**

Here you can copy items from one questionnaire to another

- [Copy Item Between Questionnaires](#)

**Questionnaire Fonts and Appearance**

Here you can change the fonts used in your questionnaires and make some adjustments to their appearance.

Please choose a questionnaire:

- [Update Fonts](#)

**Other Links**

- [Preview Translation Page](#)

### ***QAT Questionnaire Editing Features***

When users open the QAT Editor, they are presented with a view of the structure of an entire questionnaire. It is important to note, though, that this is not the view that respondents will see during the Field Trial and Main Study phases of PISA 2022, this tool is used to define the elements of the questions that will be displayed to respondents through a runtime. Figure 20.2 and Figure 20.3 show the main view of the QAT for a National Project Manager (NPM).

Figure 20.2. Questionnaire Authoring Tool: Main View (with a specific question example)

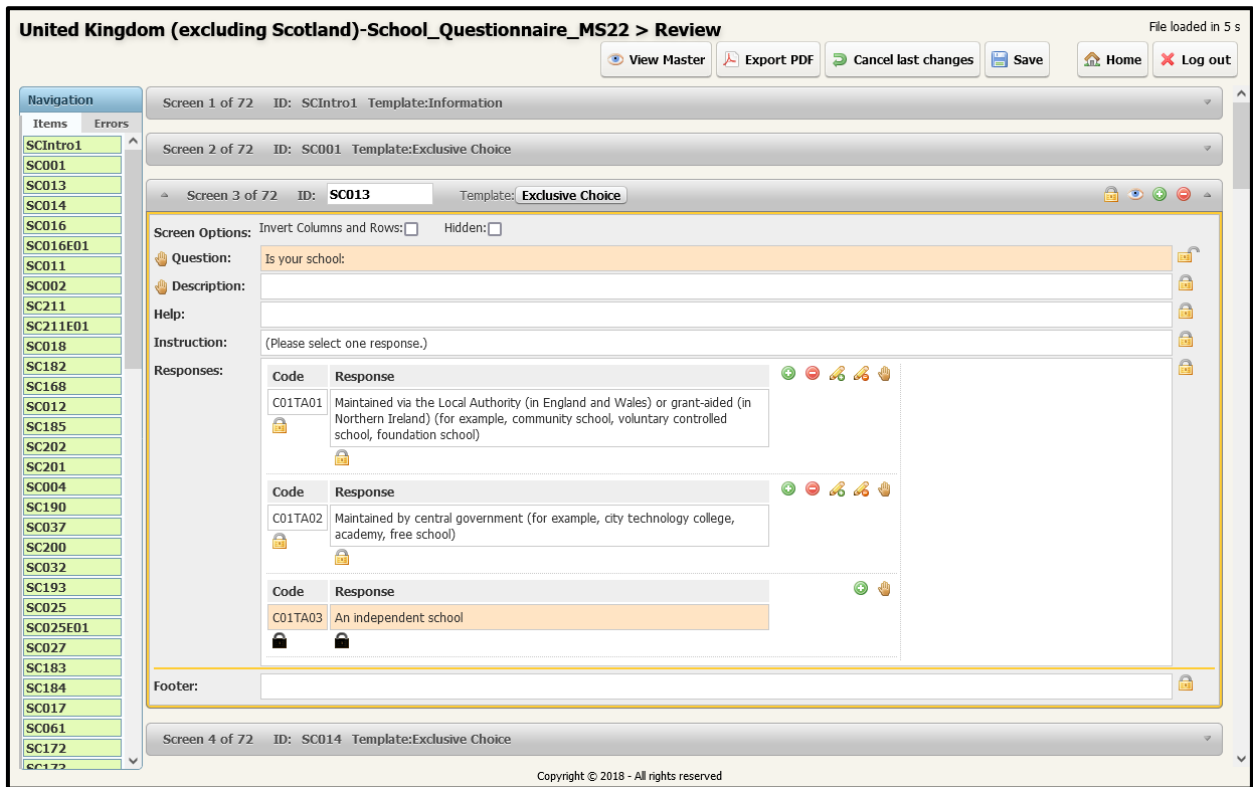
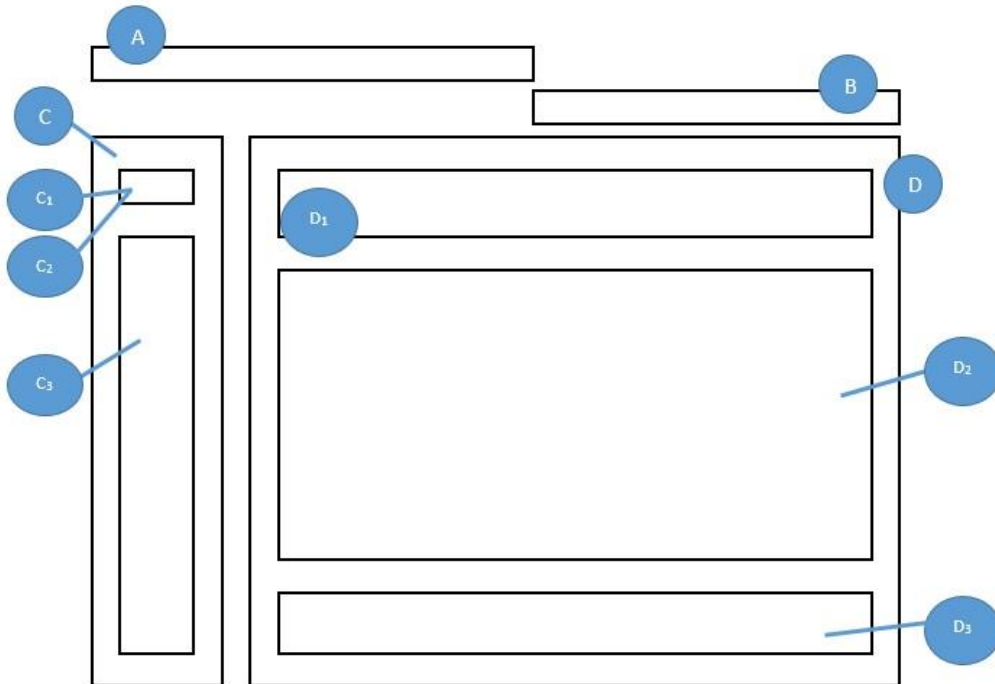


Figure 20.3. Questionnaire Authoring Tool: Organisation of Main View



The organisation of the main view is the following:

Panel A: The Questionnaire Title contains the questionnaire label (country and type of questionnaire) and the questionnaire mode (i.e. the modes of the QAT are important to note as they define the rights of a current user. Depending on the mode, the access for modifying questionnaires in the QAT editor is locked or unlocked, allowing users to work independently).

Panel B. The Questionnaire Toolbar provides the following options:

- View Master – Opens the Master English version of the current questionnaire.
- Export PDF – Generates a PDF file of the current version of the questionnaire.
- Cancel Last Changes – Undoes any changes since the last time the user has saved their work on the questionnaire.
- Save – Saves the questionnaire to the database. When clicked, this action also provides a check for whether routing rules and consistency checks are correctly formatted, in the questionnaire. If the test fails, the user will receive a notification that there are currently errors in the questionnaire.
- Home – Redirects the user to the QAT homepage.
- Log Out – Disconnects the user from the QAT platform.

Panel C: The Navigation Panel lists contains the following elements:

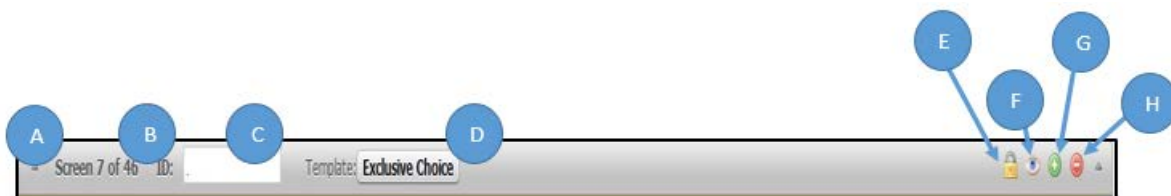
- The questionnaire items (C<sub>1</sub>) or
- A list of errors currently present in the questionnaire (C<sub>2</sub>)
- Quick access to questionnaire screens (C<sub>3</sub>)

Panel D: The QAT Editor displays the list of all questions (referred to as “screens”) and rules (referred to as “rules headers”) available for a questionnaire. When clicked, each part will expand or collapse a specific screen or rule window (D, D<sub>1</sub>, D<sub>2</sub>, D<sub>3</sub>).

### *Questions Expanded View*

When a specific screen is expanded in the QAT editor, additional features are available. Figure 20.4 shows the Expanded View information. Inside the expanded view, the user can edit the different parts of a questionnaire screen using the QAT editor: the question text, description, instruction, help, and response categories/options.

**Figure 20.4. The expanded view information**



The features available for users in the questionnaire screens include:

- A. Show/Hide Screen** button can expand or collapse a specific questionnaire screen or rule header.
- B. Screen Number** label shows the location of the screen in the sequence of questionnaire items out of the total number of questionnaire screens.
- C. Screen ID** displays the technical identifier of the screen and rule headers (i.e. SC025).

**D. Template** label displays the name of the template used for editing the questionnaire screen (see section about questionnaire templates for additional information).

**E. Lock/Unlock** button makes a questionnaire screen editable or not for a National Project Manager. This button is not available to NPMs.

**F. Preview** button opens a preview of an item, giving the user a view of how the question stem, response options, helps and instructions will be displayed.

**G. Add Screen** button inserts a new question or rule in the questionnaire just below the currently expanded item.

**H. Delete Screen** button will remove the question or rule from the questionnaire. Users who click this button will first receive a notification asking for confirmation of deletion.

### Previewing Questionnaire Items

The questionnaire platform offers three preview options for reviewing and checking the quality of the questionnaires. The first option is a question preview panel that can be accessed from within the QAT Editor using the Preview button available in the expanded view of each question. In this preview mode users see only the screen for the individual item selected *with the English source text*. This preview tool is helpful for reviewing the general layout of the question and the IDs for each response field to better understand how data will be labelled. This preview tool is shown in Figure 20.5.

Figure 20.5. Preview of a question in the QAT

PISA 2022 List of items

**Is your school a public or a private school?**

(Please select one response.)

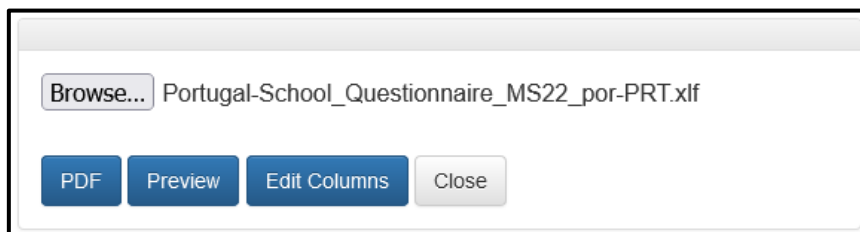
A public school (This is a school managed directly or indirectly by a public education authority, government agency, or governing board appointed by government or elected by public franchise.)	SC013Q01TA01 <input type="radio"/>
A private school (This is a school managed directly or indirectly by a non-government organisation; e.g. a church, trade union, business, or other private institution.)	SC013Q01TA02 <input type="radio"/>

Reset ◀ ▶

The second option is to preview the full national adapted questionnaire in the English source version using the “Review and Test the Questionnaire” link on the QAT homepage. This option lets users navigate through the entire questionnaire in a test environment to confirm the agreed-upon adaptations and routing are working appropriately before beginning translation.

The third option is to preview the national questionnaire in the language of administration. This tool integrates the translation file that the country has worked on separately into the questionnaire runtime. This allows users to see the questionnaire as it will be administered to students, school administrators, or teachers. Users access this preview by clicking “*Upload XLIFF for Preview*” from the home page of the QAT and then upload their translation file (XLIFF format) as shown in Figure 20.6.

**Figure 20.6. Upload XLIFF for Preview feature in the QAT**



The Upload XLIFF for Preview feature allows NPMs to view their translated questionnaire materials in a runtime environment identical to what questionnaire respondents would see in either the Student Delivery System (SDS) or in the online questionnaire. NPMs may also use create a translated PDF version of their questionnaires.

The XLIFF previewer an “Edit Columns” tool that allows users to adjust the width of response columns for each question and language version individually to ensure that translated text is not truncated. The features of this tool are shown in Figure 20.7. For PISA 2022, this tool was used only by PISA Administrators.

Figure 20.7. Upload XLIFF for Preview – Edit Columns feature

United Kingdom (excluding Scotland)-School\_Questionnaire\_MS22

Select an item▼

Load English (United Kingdom (excluding Scotland)) eng-ZZZ (Read Only)
  Space Response Columns Evenly

Load English (United Kingdom (excluding Scotland)) eng-QUK

Save English (United Kingdom (excluding Scotland)) eng-QUK layout

---

**Which of the following definitions best describes the community in which your school is located?**

*(Please select one response.)*

A village, hamlet or rural area (fewer than 3 000 people)	SC001Q01TA01 <input type="radio"/>
A small town (3 000 to about 15 000 people)	SC001Q01TA02 <input type="radio"/>
A town (15 000 to about 100 000 people)	SC001Q01TA03 <input type="radio"/>
A city (100 000 to about 1 000 000 people)	SC001Q01TA04 <input type="radio"/>
A large city (with over 1 000 000 people)	SC001Q01TA05 <input type="radio"/>

SC001

◀ ▶

### Question templates

The QAT editor is a template-based questionnaire authoring system that supports the creation of multilingual content (this includes left-to-right and right-to-left texts, and 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 consistency checks. In PISA 2018, national centres entered their translations directly into the QAT and so the support of the languages within the system was important; however, in PISA 2022 the national centres entered the English back-translation of their agreed-upon content adaptations in the QAT and then the QAT generated electronic translation files (XLIFFs) that could be used by standard computer-assisted translation tools.

All PISA 2022 questions were authored using one of the following screen templates available through the QAT editor:

- Drop Down (Table)
- Drop Down
- Exclusive Choice
- Multiple Choice
- List of Text Inputs
- Free Text Input
- List of Exclusive Choice (Table)
- List of Multiple Choice (Table)

- Multiple List of Text Inputs (Table)
- Scale Question Type
- Information

Additionally, there were two templates for defining rules that were used within the questionnaires:

- Consistency Check Rule
- Routing Rule

A short description of each template is provided below, with examples in Figure 20.8 through Figure 20.21.

### Figure 20.8. Information Template

PISA 2022 List of items

Dear <school administrator>,  
 Thank you for participating in this study. This questionnaire asks for information about:

- School background information
- School management
- Teaching staff
- Assessment and evaluation
- Targeted groups
- School climate

This information will help illustrate the similarities and differences between groups of schools in order to better establish the context for students' test results. For example, the information provided may help to establish what effect the availability of resources may have on student achievement – both within and between countries.  
 The questionnaire should be completed by the principal or designate. It should take about 45 minutes to complete.  
**For some questions specific expertise may be needed. You may consult experts to help you answer these questions.**  
 If you do not know an answer precisely, your best estimate will be adequate for the purpose of the study.

Please note that the forward button used to proceed to the next question is located at the bottom right hand corner of your screen. In some instances you may need to scroll down to the bottom of your screen to access this forward button.

**Your answers will be kept confidential. They will be combined with answers from other principals to calculate totals and averages in which no school can be identified.**

<School reminder note>

Reset ◀ ▶



The *Information* template shown in Figure 20.8 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), introduce a new section of questions, and to thank the respondent at the end of the questionnaire for their participation.

The *Exclusive Choice* template shown in Figure 20.9 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 is a pre-assigned response number assigned to each radio button (e.g. values 01, 02, 03, 04, 05, or 06 is assigned to each of the radio buttons shown in Figure 20.9). 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 20.9. Exclusive Choice Template

PISA 2022



List of items  

**Which of the following definitions best describes the community in which your school is located?**

*(Please select one response.)*

A village, hamlet or rural area (fewer than 3 000 people)	<input type="radio"/>
A small town (3 000 to about 15 000 people)	<input type="radio"/>
A town (15 000 to about 100 000 people)	<input type="radio"/>
A city (100 000 to about 1 000 000 people)	<input type="radio"/>
A large city (1 000 000 to about 10 000 000 people)	<input type="radio"/>
A megacity (with over 10 000 000 people)	<input type="radio"/>


SC001

Reset  

The *Multiple-Choice* template shown in Figure 20.10 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 when a user clicks on them and unselects if clicked a second time.

Figure 20.10. Multiple-Choice Template

PISA 2022

List of items 

**This school year, which types of <additional mathematics instruction> do you participate in?**

*(Please select all that apply.)*

One-on-one tutoring with a person	<input type="checkbox"/>
Internet or computer tutoring with a programme or application	<input type="checkbox"/>
Video-recorded instruction by a person	<input type="checkbox"/>
Small group study or practice (2 to 7 students)	<input type="checkbox"/>
Large group study or practice (8 or more students)	<input type="checkbox"/>
I do not participate in <additional mathematics instruction>	<input type="checkbox"/>

ST297




Reset  

Figure 20.11. List of Exclusive Choice (Table) Template

PISA 2022

List of items 



**The following questions concern your home. If you live in multiple homes, please consider the <home> you spend most of your time in.**

**Which of the following are in your <home>?**

*(Please select one response in each row.)*

	Yes	No
A room of your own	<input type="radio"/>	<input type="radio"/>
A computer (laptop, desktop, or tablet) that you can use for school work	<input type="radio"/>	<input type="radio"/>
Educational Software or Apps	<input type="radio"/>	<input type="radio"/>
Your own <cell phone> with Internet access (e.g. smartphone)	<input type="radio"/>	<input type="radio"/>
Internet access (e.g. Wi-fi) (excluding through smartphones)	<input type="radio"/>	<input type="radio"/>
<country-specific>	<input type="radio"/>	<input type="radio"/>
<country-specific>	<input type="radio"/>	<input type="radio"/>

ST250

Reset  

The *List of Exclusive Choice (Table)* template shown in Figure 20.11 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 item, and the columns are the response options for each item. In addition, the QAT editor allows the author to invert the table, so that items are in the columns and the response options are in the rows. Typically, this template presents a single question text or stem in the blue box at the top of the screen, and the items that are part of that question are represented in each row.

The *List of Multiple Choice (Table)* template shown in Figure 20.12 presents the respondent with one or more non-exclusive choice questions on a single screen in a tabular format. It is like the previous template; however, it uses checkboxes so that more than one choice can be selected for each item (row), or column if the presentation is inverted. The data generated by this screen include a response of 0 (unchecked) or 1 (checked) for each response option for each question. In the example shown in Figure 20.15 the screen will generate 12 individual variables of data.

**Figure 20.12. List of Multiple Choice (Table) Template**

PISA 2022 List of items

**Who usually lives at your <homes> with you?**

*"Main <home>" refers to the home where you spend most of your time.  
(Please select all that apply in each column.)*

	At my main <home>	At my other <home(s)>
Mother or other female guardian	<input type="checkbox"/>	<input type="checkbox"/>
Father or other male guardian	<input type="checkbox"/>	<input type="checkbox"/>
Brother(s) (including stepbrothers)	<input type="checkbox"/>	<input type="checkbox"/>
Sister(s) (including stepsisters)	<input type="checkbox"/>	<input type="checkbox"/>
Grandparent(s)	<input type="checkbox"/>	<input type="checkbox"/>
Other relatives (e.g. aunt, uncle, cousin)	<input type="checkbox"/>	<input type="checkbox"/>



ST229

Reset ◀ ▶

The *List of Text Inputs* template shown in Figure 20.13 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 of a minimum/maximum numeric value or text length can be placed on the values entered in each case.

Figure 20.13. List of Text Inputs Template

PISA 2022

List of items  

**As of <February 1, 2022>, what was the total school enrolment (number of students)?**

*(Please enter a number for each response. Enter "0" (zero) if there are none.)*

Number of boys:	<input type="text"/>
Number of girls:	<input type="text"/>

SC002





Reset  

Figure 20.14. Multiple List of Text Inputs (Table) Template

PISA 2022

List of items  



**How many of the following teachers are on the staff of your school?**

**Include both full-time and part-time teachers.** A full-time teacher is employed at least 90% of the time as a teacher for the full school year. All other teachers should be considered part-time. Regarding the qualification level, please refer only to the teacher's **highest qualification level.**

*(Please enter a number in each space provided. Enter "0" (zero) if there are none.)*

	Full-time	Part-time
Teachers in TOTAL	<input type="text"/>	<input type="text"/>
Teachers <fully certified> by <the appropriate authority>	<input type="text"/>	<input type="text"/>
Teachers with an <ISCED Level 6 - Bachelor's or equivalent level> qualification	<input type="text"/>	<input type="text"/>
Teachers with an <ISCED Level 7 - Master's or equivalent level> qualification	<input type="text"/>	<input type="text"/>
Teachers with an <ISCED Level 8 - Doctoral or equivalent level> qualification	<input type="text"/>	<input type="text"/>

SC018

Reset  

The *Multiple List of Text Inputs (Table)* template, shown in Figure 20.14, 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. Like the previous template, the response values can be either text or numeric, and can be limited in their range.

The *Scale Question Type (slider)* template shown in Figure 20.15 is used to collect numeric information on a sliding scale. 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. Step values for the sliders can be set, and the author may include labels for the left and right ends of the scale. The slider differentiates between no response (not moving the slider at all) and moving the slider to the “0” position. PISA 2022 did not use the scale question type template for new questions, but there were a handful of trend questions that still used this template this cycle.

**Figure 20.15. Scale Question Type Template**

PISA 2022 List of items

**During the last three months, what percentage of teaching staff in your school has attended a programme of professional development?**

*A programme of professional development here is a formal programme designed to enhance teaching skills or pedagogical practices. It may or may not lead to a recognised qualification. The programme must last for at least one day in total and have a focus on teaching and education.*

*(Please move the slider to the appropriate percentage. If none of your teachers participated in any professional development activities select "0" (zero).)*

All teaching staff at your school  0% 100%

Staff who teach mathematics at your school  0% 100%

SC025 Reset

The *Free Text Input* template shown in Figure 20.16 supports an open-ended text response. The respondent is presented with a large text box in which they can enter a long response with line breaks to provide multiple paragraphs. This template was only used in national questions in PISA 2022.

Figure 20.16. Free Text Input Template

PISA 2022 List of items

**Please describe your Mother's main job:**

SC801

Reset ◀ ▶

Figure 20.17. Drop-Down Template

PISA 2022  [List of items](#)

**How old were you when you arrived in <country of test>?**

*(Please select from the drop-down menu to answer the question. If you were less than 12 months old, please select "age 0-1" (age zero to one).)*

ST021




[Reset](#)

The *Drop-Down* template shown in Figure 20.17 presents the respondent with one or more drop down menus from which to select their response to a question. Each menu can have a textual label to present a question or to indicate what type of information (e.g. age) the respondent should select from the menu. The contents of a menu are defined using a list with each text response in the list assigned a number in the QAT editor. The menus can share the same list of response values across items, or each item on the screen can have a unique list.

Like the *Drop-Down* template, the *Drop-Down (Table)* template shown in Figure 20.18 presents the respondent with one or more drop down menus for providing a response. In this template, the menus are organised into a table. The drop-down menu contents themselves 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.

Figure 20.18. Drop-Down (Table) Template























PISA 2022

List of items ?   



**In your school, are <standardised tests> and/or teacher-developed tests of students in <national modal grade for 15-year-olds> used for any of the following purposes?**

*If you need further explanation of the term "<standardised tests>", please use the help button.*

*(Please select either "yes" or "no" to indicate the use of <standardised tests> and teacher-developed tests for each of the specified purposes.)*

	<Standardised tests>	Teacher-developed tests
To guide students' learning	Select... 	Select... 
To inform parents or guardians about their child's progress	Select... 	Select... 
To make decisions about students' retention or promotion	Select... 	Select... 
To group students for instructional purposes	Select... 	Select... 
To compare the school to <district or national> performance	Select... 	Select... 
To monitor the school's progress from year to year	Select... 	Select... 
To make judgements about teachers' effectiveness	Select... 	Select... 
To identify aspects of instruction or the curriculum that could be improved	Select... 	Select... 
To adapt teaching to the students' needs	Select... 	Select... 
To compare the school with other schools	Select... 	Select... 
To award certificates to students	Select... 	Select... 

SC035

Reset  

### 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 condition (i.e. "True" or "False") intended to represent the logic of the rule that checks the values of some response variables from different questions the respondent has answered. If the condition evaluates "True," a notification message is displayed to the user. The template for defining the consistency check rule appears in Figure 20.19.

Figure 20.19. Consistency Check Rule Template

Rules header ID: SC164E01 Template: Consistency rule

Consistency Check

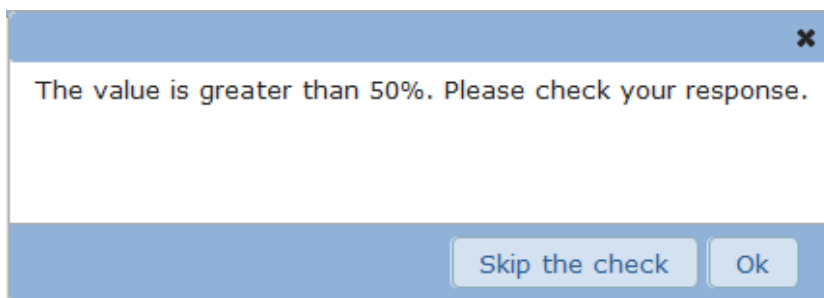
Rule: IF ^SC164Q01HA01 > 50

Message: The value is greater than 50%. Please check your response.

The rule is evaluated when the respondent navigates away from the current question, by clicking either Forward, Back, or Log Out. When the condition is true, a message is shown like the one in Figure 20.20.



Figure 20.20. Consistency Check Message



The consistency check is a soft check and will not require the respondent to change their answer if the check appears. The respondent can click on “OK” in the check and go back to the current question to change their response. If the respondent clicks the “Skip the Check” button, the questionnaire will proceed as normal.

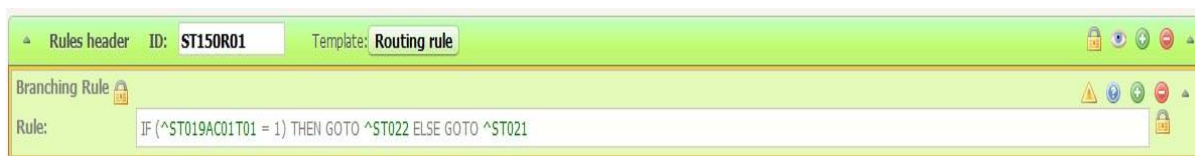
### **Routing Rule**

The *Routing Rule* template allows the author to use branching within a questionnaire to direct the question flow. Routing rules appear in between questions in the questionnaire, and they are executed after the completion of the question before the rule.

The routing rules are based on specific conditions, like the consistency checks. The rules are defined using IF—THEN--ELSE logic. If the condition evaluates “True” the “Then” portion is executed, otherwise the “Else” part is executed. The “Then” and “Else” parts can be either another IF--THEN--ELSE rule 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. For example, an open response where they can type their answer. In PISA 2022 field trial routing rules were used for the first time to create virtual booklets to indicate that certain questions should be skipped if a student’s random number was within a certain range. An example of a routing rule can be seen below in Figure 20.21.

Figure 20.21. Routing Rule Template



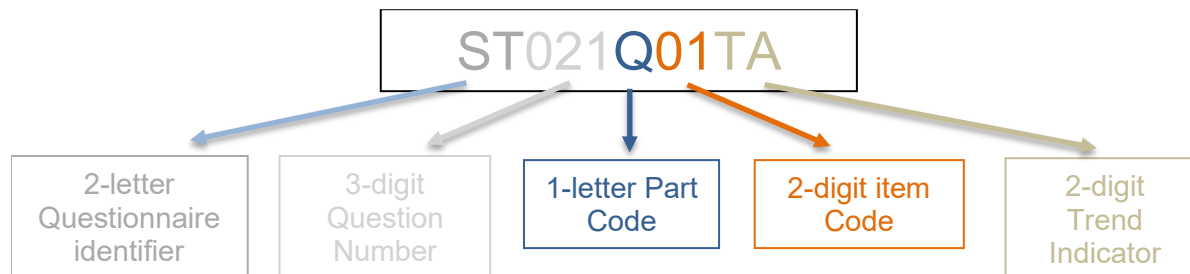
### *Identifiers within the QAT*

An identifier (or ID) is a tag attached to an object in the QAT. When authoring questionnaires, it is important that each question, item, and rule has an ID that follows a standard convention so that each object in the questionnaire can be appropriately identified and data from the questionnaires is generated in a standard format.

In the QAT editor, the types of objects receiving an ID are the various questions, helps, instructions, and response options. These IDs and tags are used when importing the translations used to display the

questionnaires in each local language. IDs are also used for each rule and for each element designed to receive and store the data provided by the respondents (i.e. answers). IDs used for data capture within the questionnaires are at least 10 characters long and follow a set format shown in Figure 20.22

**Figure 20.22. Question IDs**



The interpretation of these IDs is as follows:

- The 2-letter questionnaire identifier indicates the questionnaire in which the item was administered. ST for Student Questionnaire or Student Questionnaire-UH, SC for School Questionnaire, FL for Financial Literacy Questionnaire, IC for ICT Familiarity Questionnaire, WB for WBQ, and PA for Parent Questionnaire.
- The 3-digit question number: this is a unique ID given to the particular screen on which the questions are administered and can be used for either a single question or a set of questions presented in a table on the screen. As much as possible, questions retain their numbers across cycles to allow for easier identification of trend variables. Question numbers beginning with 800 indicate a national question administered only in a particular country.
- The 1-letter part code introduces the item code and indicates whether the question is equivalent to the master (a code of Q) or is a country-adapted variable requiring harmonization to be compared to the master (a code of C). Consistency checks are labelled with the code E, and routing rules are labelled with the part code R.
- The 2-digit item code indicates the number of the individual question item administered on the screen. If a screen (such as a table screen) contains four items, these numbers will typically range from 01 to 04; however, due to trend IDs or elimination of items after the field trial, the item code is not always sequential on a screen.
- The 2-digit trend indicator is used to indicate the cycle in which the question was originally introduced in order to facilitate trend analysis. The trend indicators are shown Annex Table 20.A.3

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 a question) with an unexpected or inappropriate ID is unusable and can eventually not be analysed. Checking the consistency of IDs was one of the critical tasks performed by contractors when authoring and reviewing the computer-based questionnaires.

## General questionnaire development process

The life-cycle of a questionnaire in PISA followed a process that can be split in ten major steps. These steps are described in Figure 20.23.

This sequence of steps took place twice: once for the Field Trial (FT) and again in an abbreviated process for the Main Survey (MS). During the Field Trial, the whole platform (i.e. the tools, computer servers,

network access, etc.) and the material (i.e. the questionnaires) were tested with a limited sample of respondents. After the Field Trial, the results and feedback collected are analysed and reviewed. Then, for the Main Survey, the sequence was started for a second time and each step integrated all necessary adjustments in terms of process, questionnaires material, and tooling. This double-phase cycle provided better data quality.

The procedure for creating the paper-based questionnaires was the same, except step 2 (authoring the questionnaires in the computer platform) and step 5 (quality checks on the implementation of the adaptations in English) were omitted.

In the following sections, each step of this process is explained in more detail.

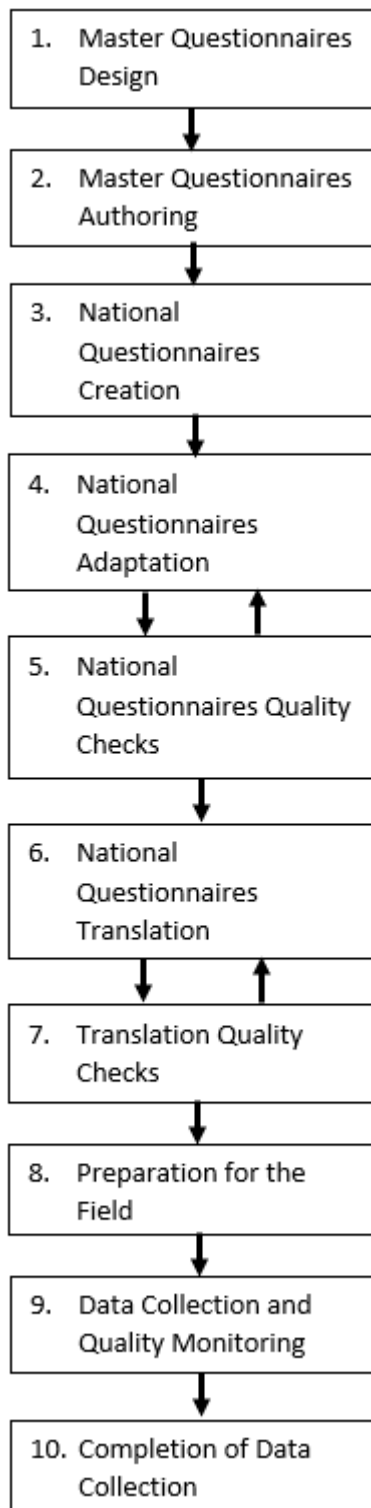
### ***Step 1: Master questionnaires design***

The master versions of the questionnaires were created as Word documents, and each contained information about the trend status, IDs, application of matrix sampling to the question, and routing instructions. Once the master versions of the questionnaires were finalized, the Word document was used to create the paper-based master version to be adapted and translated by paper-based participants. The contractors then used the master Word version to begin authoring the computer-based questionnaires in the platform.

### ***Step 2: Master questionnaires authoring***

Contractors first used the QAT to author the international master version of the questionnaires in English. Trend questions from the previous cycle were copied from the main survey PISA 2018 master questionnaire profiles to ensure across-cycle continuity of question templates, IDs, data format (e.g. string, numeric), and range limits. The appropriate question template was chosen for each of the new items and the appropriate data types, range restrictions, and consistency checks were added. Questions were ordered appropriately to follow the design-specified routing through the questionnaire and routing rules were inserted. Each questionnaire's formatting was reviewed to ensure it had the appropriate layout, and the master version of the questionnaires were tested extensively using testing scenarios to ensure a high-quality initial version.

Figure 20.23. PISA 2022 computer-based questionnaire life cycle



Step 1: The **Master Questionnaires were designed** in collaboration with the Questionnaire Expert Group. These questionnaires are first created as Microsoft Word documents that will become the Master Paper-Based Questionnaires.

Step 2: The computer-based versions of the **Master questionnaires were authored** using a unique authoring tool in the PISA questionnaire platform. They were produced in English then reviewed and tested.

Step 3: The Master Questionnaires were duplicated for the participating countries. These questionnaires, called **National Questionnaires, were created** and made available to countries for adaptation.

Step 4: The **adaptation of National Questionnaires** was performed by members of the national centres. The adaptations take the form of adding or suppressing questions or changing parts of questions as required by the national context.

Step 5: The **quality of the adapted National Questionnaires was checked** against the original Master Questionnaire. The quality of adaptation is important for guaranteeing that the collected results are comparable at the international level.

Step 6: The National Centre **translated the questionnaire text** into each national language version administered.

Step 7: The **quality of the translations of the National Questionnaires** was checked against the agreed-upon national adaptations.

Step 8: When a questionnaire had successfully passed all the quality and technical checks, it was **prepared for the field**. The deployment was either online (via a connection to Internet) or as part of the Student Delivery System.

Step 9: During the data collection periods, **data was collected** either online or in the schools, depending on the distribution method of the questionnaires.

Step 10: At the **end of the data collection**, the online National Questionnaires were deactivated, and respondents could no longer access them. Final data files were exported for data cleaning and analysis.

### **Step 3: Creation of national questionnaires**

Once the master questionnaires were authored and finalized, they were used as the template to create the national questionnaires for each country. The contractors duplicated the master questionnaire for each country, so every participant started with the same set of questions in the same order. In order to maintain trend adaptations from the previous cycle, trend question screens were copied from the country's PISA 2018 main survey questionnaire instead of the master questionnaire. Since PISA 2018 questionnaires were translated in the profile, as part of this copy process, the English back-translation of the PISA 2018 adaptation was inserted into the QAT editor. The initial version of each national questionnaire contained a combination of new questions for PISA 2022 copied directly from the master questionnaire and trend questions from PISA 2018 copied from the country's final PISA 2018 computer-based Main Survey Questionnaire. This copy operation was performed by the contractors using several system scripts. These national questionnaires were then put into a mode that allowed the national centres to adapt the content.

### **Step 4: National questionnaire adaptation**

At this step in the process, the National Centre first documented in a spreadsheet all the structural adaptations needed to the questionnaires, including adding or deleting questions and response options and all required content adaptations such as the specific names of study programmes. The contractors reviewed and approved the adaptations to ensure internationally comparable questionnaires. Once all adaptations were negotiated, the National Centre connected to the QAT to view and edit their national questionnaires in the platform and insert the agreed-upon adaptations. All adaptations were inserted into the QAT in English so that the text in the national questionnaire in the QAT became the nationally adapted English source text used later for translation. Much like for authoring the master questionnaires, the National Centre had 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.

When opening the questionnaire in the QAT, the National Centre could see and edit the questions for the new PISA 2022 content. Trend questions copied from the PISA 2018 cycle were locked so that the National Centre could not edit them, and any changes approved for these questions were implemented centrally by the contractors. Maintaining the quality and integrity of the trend questions over time is important to be able to analyse data across cycles.

### **Step 5: National questionnaire Quality Check**

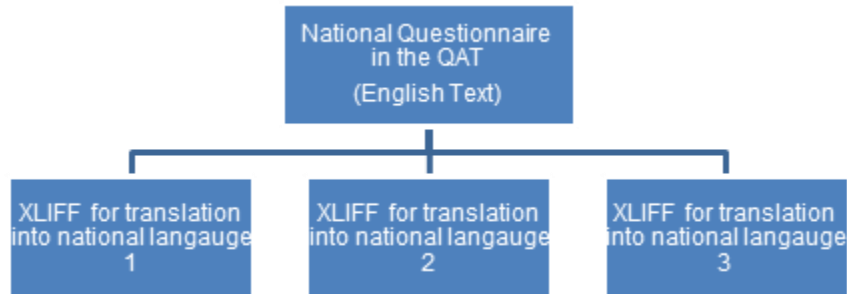
Once a country's adaptations and national questions were implemented in the QAT, the national centres tested the questionnaire using contractor-prepared testing scenarios to review all adaptation in English and confirm the routing of the questionnaire worked as expected. Then the contractors reviewed and approved the national version in the QAT to confirm all the agreed-upon adaptations had been correctly authored and to centrally insert any agreed-upon changes to trend adaptations. The contractor also carefully reviewed the questions to ensure all agreed national questions had been inserted in the questionnaire and reviewed questionnaire IDs to confirm that they were appropriately updated to conform to the ID conventions for the cycle. IDs are the key identification point for the data analysis and an error in this part might result in loss of data. After these quality checks, the questionnaires were locked in the QAT so that no further edits could be introduced by the national centre.

### **Step 6: National questionnaire Translation**

With the national questionnaire adaptations finalized, the contractors then used the national questionnaire in the QAT to generate the country-specific English source XLIFF files for translation. Each country had one single structure and source text per questionnaire that then could be translated into multiple languages (see Figure 20.24). The XLIFF files were inserted into an OmegaT translation project which allowed the

National Centre to reference translations of similar text used in PISA 2018 to speed up translation. The OmegaT project also allowed the translation of items from PISA 2018 to be automatically filled with the trend translation and locked so the National Centre could not edit them.

**Figure 20.24. Translation of Questionnaires into multiple national languages**



Countries performed double translation of new questionnaire items and reconciled those translations. If updates were needed to trend translations, the country documented those changes for review by the contractors. During the translation process, countries could generate translated XLIFF files and upload them to the preview tool in the QAT to review the questionnaire as a respondent would. As part of the preview process, countries also noted any layout issues that needed to be fixed by the contractors.

### **Step 7: National questionnaires quality check**

After the National Centre completed the translation of the questionnaires, the contractors checked the quality of the translations, the accuracy of the translation compared to the English master version, the routing and formatting, and the IDs and data generation of the questionnaires.

National translations were reviewed by verifiers under the direction of the translation contractors to confirm that all adaptations were appropriately translated, all translation notes from the questionnaire developers were followed, and to confirm the accuracy of any requests to update translations of trend items. Verifiers updated translations or provided notes to national translators to address issues as necessary. The translation and adaptation discrepancies were documented in a spreadsheet which was delivered to the National Centre for their review. The National Centre was able to accept or refuse these comments and could update their translations accordingly.

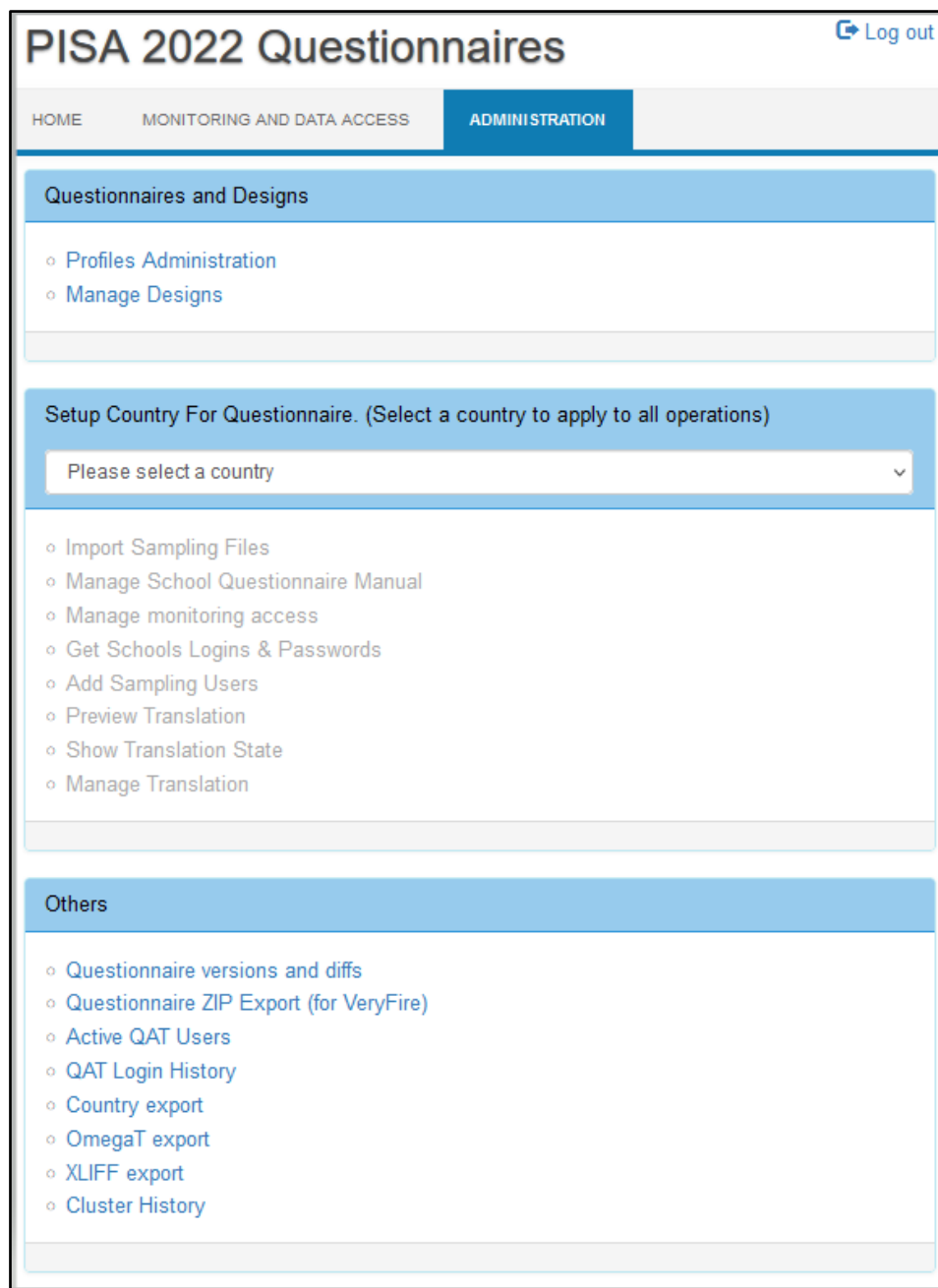
The contractors also reviewed the questionnaires to manually check that a user was able to go through the questionnaire from the beginning until the end without a software error due to, for instance, errors in routing rules; check if all questions and messages were translated; and check if all the parts of the interface were translated and well-integrated.

National centres were provided testing scenarios for each questionnaire to validate the accuracy of their translation and adaptation work. These testing scenarios defined different ways in which a respondent could answer a questionnaire following every possible routing. National centres were required to test the questionnaires in each language version following these scenarios and provide their test results to the technical team for review. The contractors reviewed the output files to confirm no technical problems were detected when saving data. Only once all these reviews were completed were questionnaires deemed ready for administration.

### Step 8: Preparation of national questionnaires for delivery

To prepare the questionnaires for administration, the QAT administrators and technical team used the features of the questionnaire platform's administrative interface shown in Figure 20.25.

Figure 20.25. Questionnaire Platform – Administrative View



There were two modes of delivery used for the questionnaires in PISA 2022. The student questionnaires, including the optional ICT, Financial Literacy, and Well-Being questionnaires, were run as part of the PISA student delivery system (SDS). The School and Teacher questionnaires were delivered online over the

Internet. Both delivery modes shared a common code base and database structure, but the preparation for delivery followed different procedures.

For the student questionnaires, the preparation step involved uploading the final translation file for each national language version into the QAT and then exporting the completed national questionnaires for each country, as well as the questionnaire software and user interface translations, in a form that could be integrated into the Student Delivery System (SDS) and loaded onto and accessed from USB drives. The export only included the software components needed to run the questionnaire, so components such as the QAT, were not included in the export, and a database image with the national questionnaires was created. These exported files were directly integrated into the PISA SDS software for a country, and then tested and validated.

The online School and Teacher questionnaires required more steps to prepare for delivery. The first critical step in this process was to import the sampling information from each computer-based testing country into the questionnaire platform so that the selected schools and teachers would be known to the system and could be identified when they connected to complete the questionnaires online. To do this, the final approved “Sampling Task 5b” (specific to the Field Trial) or “Sampling Task 11” (specific to the Main Study) output files were taken from the PISA Portal and uploaded into the QAT. The content of these forms is described in Chapter 6 of this Technical Report. These files contained the list of schools selected from the sampling process, using anonymous ID codes. The QAT software used these files to generate logins and passwords for each sampled school. These logins and passwords were then sent to the national centre, which distributed them to the selected schools and teachers accordingly. The ACER Maple sampling software generated the IDs and passwords used for the Teacher Questionnaire, and Teacher Questionnaire authentication process was set up to recognize valid teacher IDs and passwords.

The countries participating in the online questionnaires in PISA 2022 were spread out across the world. To improve performance for end users, servers were set up on several continents, as shown below in Figure 20.26, following the same distribution used in the 2018 PISA cycle. The distribution of servers helped to reduce network latency and improved the performance. Server installations for this cycle of PISA were in Germany, Singapore, Australia, and the United States.



Figure 20.26. Distribution of the PISA 2022 servers



Participating countries/economies were routed to their nearest server locations. When respondents logged into the questionnaire using their user ID and password on the School or Teacher login website, they were automatically redirected to their assigned server based on their login ID to complete the questionnaire. 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, and were not rerouted through the central PISA server.

### **Step 9: Data collection and quality monitoring**

During the field trial and main survey data collection periods, students, school principals, and teachers responded to the questionnaires. For the students, responses were captured as part of the PISA Student Delivery System, which ran from either a USB drive on a school computer, from laptops with the SDS software pre-loaded, or through Google Chromebooks. The system ran in full screen, locked down mode when running on Windows and Macintosh computers, and in “kiosk mode” (<https://chromeos.dev/en/kiosk>) on Chromebooks.

The questionnaire software ran offline, in a standalone mode on the school computer, and all results were saved back to the USB drive. The students did not need to login to start the questionnaire. Identification and authorization of the students was performed by the Student Delivery System.

For the online questionnaires for school principals and teachers, delivery was performed online over the Internet. Schools were assigned login IDs and passwords as part of the sampling process in Step 6. A set number of teacher questionnaire IDs were accepted for each school. When respondents first connected to the questionnaire platform, they entered their ID and password. The questionnaire software selected the appropriate national questionnaire based on this ID. In countries with multiple language versions of their questionnaires, users had to select which language they wanted to use before proceeding further.

As respondents completed the questionnaires, data was collected by the questionnaire platform. The original data saved was the response to each question or item. This data depended on the template used

for each question. For questions that used radio buttons, the data value saved was the response ID associated with that radio button. For checkboxes, a data value was saved for each of these boxes on the screen and the value would be zero or one depending on whether the box was selected. For sliders, drop-down menus, and textual responses, the value selected or entered is saved. If no response is selected or entered, a value of “null” is saved. For questions where matrix sampling was applied, those items that were not presented to the student received a special code so that nonresponse could be distinguished from not administered.

Along with the response data, the questionnaire saved the final valid path taken by the respondent in the questionnaire. This allowed the contractors to easily identify which questions were presented to the respondent based on the routings so that not administered questions could be distinguished from not answered. Also, a log of actions by the respondent and the questionnaire system was saved. This log includes events such as those shown in Annex Table 20.A.4.

During data collection for the online questionnaires, National Project Managers and administrators of the questionnaire platform could monitor the activity of the questionnaire respondents. The monitoring showed which respondents had connected to the questionnaire platform and how far they had progressed through the questionnaire. The platform also supported generating a PDF file for a respondent showing the questionnaire including all the responses that had been saved. The overall status for each of the questionnaires could be exported to a spreadsheet for further sorting and filtering.

During the Main Study, the sampling process selected schools to participate in the PISA survey, along with replacement schools if the originally samples schools refused or were unable to participate. Through the monitoring tools available in the questionnaire platform, the NPMs were able to activate or disable school logins to control access to the questionnaire depending on the school’s status as selected or replacement.

The administrators of the questionnaire platform had additional tools available for monitoring the progress of the respondents. These included a view of all currently connected users, as well as a history of the logins, both successful and unsuccessful. These reports were important in supporting users who reported problems and in monitoring performance issues on the servers. Additionally, the questionnaire platform saved many different logs, which the administrators used for detecting problems and troubleshooting them. All the servers were monitored and active 24 hours a day during the entire field test and main survey administration dates.

### ***Step 10: Completion of data collection***

Access to the online access to the questionnaires closed four weeks after the country’s negotiated field trial or main study data collection period ended. Once access to the questionnaires was closed, national centres exported their results data for inclusion in their national database that they submitted to the contractors. After the questionnaires were closed, respondents who attempted to login received a message indicating that the questionnaires were currently not available and asking them to contact their National Centre for further information.

Each country’s result data was available throughout the data collection period and could be reviewed for completeness by the national centre. Once data collection was complete, the national centres were required to download the final results in a single compressed file and import it directly into the Data Management Expert system for data processing.

The access to the servers and the questionnaire software was available several weeks after the end of the data collection to allow some time for the NPMs to retrieve the data and ask the contractors questions about any issues in the data that they uncovered.

## Overview of the technical infrastructure

This section describes the technical aspects of the software and hardware used to support the PISA 2022 computer-based 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 during the course of developing and extending the platform.

The PISA Questionnaire platform is composed of two primary subsystems. One, the QAT supports authoring questionnaires and managing the many national versions of the questionnaires. The second, the QAT Runtime, implements the execution environment for the questionnaires, presenting questions to the respondents, implementing branching rules, and collecting data for later analysis. The QAT software was written primarily in PHP on the server side and JavaScript within the web browser. The QAT Runtime software was new for the PISA 2022 cycle. This subsystem was split off and built from scratch to address performance issues that arose in previous cycles, as well as to support the new Electron based PISA Student Delivery System (SDS) that was specifically developed for this cycle. The QAT Runtime was built using Node.JS and Express.JS on the server, 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 in the QAT, and MongoDB (for online questionnaires) and NEDB (for offline questionnaires) in the QAT Runtime. The questionnaire content was structured using custom XML markup. The online questionnaire servers were Linux based, using Ubuntu 20.04 LTS. The student questionnaires were delivered as part of the PISA Student Delivery System, which was based on Electron. For the Main Study online questionnaires, multiple servers were deployed using the Amazon Web Services EC2 system.

## Summary

Improvements made to the authoring and delivery of questionnaires in PISA 2022 provided several advantages over the PISA 2018 cycle. First, the updates to the QAT Runtime allowed for faster administration of the questionnaires and for significant increases in the amount of material that could be administered to students. PISA 2022 introduced the use of a random number to assign students to a path in the questionnaire, which allowed multiple version experiments to be conducted during the field trial to collect data that informed future development of PISA context questionnaire items. In addition, the field trial pilot and main study adoption of within-construct matrix sampling, allowed for greater efficiency in the instruments and an increase in the coverage of the constructs to be implemented for the first time in PISA 2022.

Second, including the English back-translation of questionnaire adaptations in the QAT instead of the translated text allowed for clearer documentation of the adaptations to ensure international comparability and higher-quality translation. Contractors and data users were able to clearly distinguish between adaptations to the content of the questions and linguistic adaptations necessary for the translation of certain terms into the local language, ultimately leading to more assurances of international comparability of the questionnaire data. Third, due to the ability to export customised translation files for each country and language version, translators and translation verifiers were able to make use of translation tools to ensure repeated text used in multiple modules across questionnaires appeared consistently, ensuring that constructs appearing in multiple questionnaires were measured using as identical an instrument as possible. Also, translators and verifiers were able to clearly see in the translation files where national translations did not match the agreed-upon customised source text, reducing misunderstandings about whether adaptations were linguistic or content-related.

## Annex 20.A. Evolution and Implementation of Questionnaire Administration in PISA 2022

Annex Table 20.A.1. Chapter 20: PISA 2022 Questionnaires and Participation Metrics

Tables	Title
Web Table 20.A.5	Participation in the PISA 2022 Main Study

StatLink  <https://stat.link/84inx7>

Annex Table 20.A.2. The PISA 2022 Questionnaires

Questionnaire	Respondent	Mode of delivery	Compulsory
Student Questionnaire	Student	Computer (SDS) and paper	Yes
Student Questionnaire – Une Heure	Student	Computer (SDS)	No
School Questionnaire	School Principal or Administrator	Computer (Online) and paper	Yes
Financial Literacy Questionnaire	Student	Computer (SDS) only	No
ICT Questionnaire	Student	Computer (SDS) only	No
Teacher Questionnaire	Teacher	Computer (Online) only	No
Parent Questionnaire	Parent of selected student	Paper only	No
Well-Being Questionnaire	Student	Computer (SDS) only	No

Annex Table 20.A.3. Trend Indicator Values in PISA Question IDs

Trend Indicator	Cycle PISA Question Introduced
TA	In multiple cycles prior to and post 2009
IA	2009
WA	2012
NA	2015
HA	2018
JA	2022

Annex Table 20.A.4. List of Logged Events

Event	Description
SESSION_START	The user starts or resumes a questionnaire.
ITEM_START	The user starts an item.
HELP	The user clicks on 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.
ANSWER_SELECTION	An answer is selected or entered.
ANSWER_UNSELECTION	The user unselects a checkbox item.
RANGE_CHECK	The answer entered triggered a range check.

Event	Description
RANGE_CHECK_FAILED	The range check warning message was shown letting the user know the permitted range of answers.
CONSISTENCY	A consistency error message is displayed.
CONSISTENCY_CANCEL	The user presses OK to return to the current screen and update their answer.
CONSISTENCY_SKIP	The consistency error is skipped and the move action proceeds.

# 21 **The PISA 2022 Computer-based Platform**

## Introduction

The PISA 2022 computer-based platform was the primary mode of assessment of student skills. While paper-and-pencil versions of the assessments remained available to participating countries/economies, development of new content to represent and measure the constructs defined in the updated assessment frameworks was done for all cognitive domains and most questionnaires in the computer-based assessments. The vast majority of countries/economies chose to implement and deliver the survey on a computer-based platform to make the most of the opportunities for reporting that this option provided. All cognitive domains were delivered via computer, including the innovative domain (Creative Thinking) and the optional Financial Literacy assessment. The Student Questionnaire, including any international options, was delivered via computer to all students who took the computer-based cognitive assessments. The computer-based assessments were delivered in over 120 different language versions across the participating countries/economies.

This chapter focuses on the functionality and technical implementation of the computer-based assessments. It also describes the functionality and technical requirements of the PISA student delivery system (SDS) and the Chromebook student delivery system (CDS) used for delivery of the PISA survey in schools. Finally, it concludes with a discussion of the open-ended coding system (OECS), used for coding of student responses to open-ended questions in the cognitive assessments that required human coding.

## Item rendering

The items for PISA 2022 were implemented using the web-based technologies HTML, CSS and JavaScript®. Modern web browsers, such as the bundled Chromium™ browser v94 used in the PISA SDS, provide a suite of features and functionalities that enable attractive presentations and facilitate engaging interactivity. At the beginning of the development work, an overall user interface was designed with a common set of elements such as navigation, help and progress indicators. Items were implemented in such a way that these common elements were shared, so that the same elements were used across all items in each language version.

PISA 2022 items are generally grouped into units consisting of one or more common stimuli and a set of items, which are also referred to as questions in this chapter. Each unit was constructed independently; with the stimulus and questions components developed first in English, then translated into French to create the two harmonized source language versions. The development was done by experienced web user-interface (UI) developers using standard HTML components and adding custom functionality via JavaScript. Each unit could be viewed on its own or grouped with other units into a test form for delivery to students as part of the assessments.

In some cases, such as the interactive mathematics and scientific literacy units, common functionalities were split out into shared programming libraries that could be reused in multiple units. For example, in the scientific literacy units the experimental data tabling and management functionality was built as a shared library. The library also managed the recording of data and supported scoring of the student's performance based on unit-specific criteria. Likewise, in reading literacy, the management and display of multiple sources in tabs was encapsulated into a shared library. In mathematics literacy the spreadsheet management functionality was built and used as a shared library.

The visual aspects of the PISA 2022 items and the automated coding of student responses were both 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. These were:

- 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 interactive mathematics and 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 translations and adaptations for over 120 different national language versions of each unit, an automated process for the integration of these translations adaptations was critical. This process commenced with the initial development of the units. The HTML files that implement the display of the unit contained only HTML mark up and the text to be shown on the screen. Layout and formatting specifications were stored separately in CSS stylesheets. The text of the units was then extracted from these HTML files and saved to a standard file format, XLIFF (<http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html>), used for computer supported translation. Once a translation was completed, the XLIFF file was injected into the original source version of the unit, resulting in HTML files with the translated text of the unit.

One of the guiding principles of the platform development was that the quality of a translation is enhanced when translators can view their translation in the context of the stimulus and 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 enter their translations directly within an interface that displays the items as they would be seen by the students, but this was not technically feasible with the current authoring method. Furthermore, the visual aspects of the items, which are tightly controlled for comparability, may distract translators from the text to be translated. A good compromise was to provide translators with an easy-to-use preview feature to view their translations as functioning items from the PISA 2022 Portal. Users were able to upload a locally prepared and saved XLIFF file, with either partial or complete translation, and in a matter of seconds be able to preview the given unit in exactly the same design and layout as a student would view and interact with it within the student delivery system. This was an important factor, particularly for the more complex and interactive units across the domains. This preview also allowed countries to test and identify potential problems with their translated units before receiving the final versions packaged within the software to be used in schools. Therefore, reported problems were fixed as early in the schedule as possible.

## School computer requirements

The goal for the PISA 2022 computer-based administration was, to the extent possible, to use the computers available in the sampled schools with no modifications or upgrades to existing hardware. The PISA 2022 system supported both Windows based and Macintosh computers and offered a Chromebook administration option as a pilot study in the field trial. All these options were also supported for the main survey. The following minimum technical requirements were established for the main survey:

	Windows	Macintosh	Chromebook
CPU Speed	1000MHz (1500 MHz Recommended)	1000MHz (1500 MHz Recommended)	N/A
Operating System	Windows 7, 8, 10 or 11	Mac OS X version 10.11 or later	Chrome OS with Google Chrome web browser version 57.0 or later
Installed Memory	1280 MB	2048 MB	N/A
Available Memory	774 MB (878 MB Recommended)	774 MB (878 MB Recommended)	N/A
Screen Resolution	1024 x 768 pixels	1024 x 768 pixels	1024 x 768 pixels
USB Transfer Rate	7.5MB/s (12MB/s Recommended)	7.5MB/s (12MB/s Recommended)	Download and Upload speed of 0.5 MB/s (2.0MB/s Recommended)

Computers with higher capabilities would obviously perform better (e.g. respond faster) when delivering the survey, but the requirement listed above were the minimum settings that would provide adequate performance.

## System diagnostic

In order to verify that the available school computers met the minimum requirements, a system diagnostics application was provided to the national PISA centres within the participating countries/economies. The System Diagnostics is a version of the delivery system without the tests and questionnaires. It was intended to be given to schools to check the compatibility of the school computers with the PISA software. It checked the computer's hardware and software setup and reports results of this check back to the user, typically the test administrator or technical support staff in the school. Additionally, the user was given the option to run a modified version of the assessment using publicly available items to verify performance.

The system diagnostics was provided to countries approximately six months prior to the start of the field trial and main study. This allowed PISA centres to review the results in advance of the data collection period with time for an alternative solution to be implemented if minimum requirements were not met. Additionally, it was recommended that test administrators run the system diagnostics on the day of the test prior to conducting the assessment.

For cases where schools did not have adequate quality or quantity of computers, PISA centres arranged for test administrators to bring laptops into schools to augment the available infrastructure. In a few cases, countries chose to administer the PISA tests in all sampled schools on external 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 2022 test delivery system, called the student delivery system or SDS (CDS for the Chromebook delivery system), integrated the PISA computer-based assessments and questionnaires for a country/economy, 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 were on a single USB drive, and results were saved back to the USB drive. The SDS could also be deployed from the computer's local hard drive or a network file server or terminal server if desired. The components which made up the SDS included the following:

- Electron framework (<https://www.electronjs.org/>)
- No SQL database engine (NeDB for the SDS version and MongoDB for the CDS version)
- Chromium™ open-source project web browser (<https://www.chromium.org/chromium-os/>).

The actual test and questionnaire content were included together with these open-source applications. The PISA 2022 test delivery system was implemented to display this content to the students and collect their responses. Using components 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 2022. This included implementation of the test flow, which assigned the designated test form and questionnaires to a student, then sequences through the test units and questionnaires in the appropriate order. It also included the functionality for collecting the survey results and exporting them when the tests were completed. The PISA test delivery system was built using Electron, an open source, cross platform framework for creating applications using Chromium and Node.JS.

The system was launched by running a single executable program written for controlling the delivery of the PISA tests. Custom builds were developed for Windows and Macintosh operating systems. 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 a local web server and in memory database, then launch a browser window to begin the process.

The Google Chromium 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. For example, it was also not possible to block the Ctrl-Alt-Delete sequence under Windows, as this required installation of a custom software driver at the system level. The 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 two sessions: Session 1 – The PISA Tests and Session 2 – The PISA Questionnaires. 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 13-digit student ID assigned by the ACER Maple software as part of the sampling process. The password was also assigned by the ACER Maple software and was a 10-digit number. The first few 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 few digits were a checksum of the three-digit test form number.

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/economies, 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, in this situation, once logged in, the student would be shown a screen asking to select a language they wanted to use for the session. The test administrator would then guide students through the login and language selection process where applicable.

An important facet of the system 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 assessment. To accomplish this, the files for rendering all test materials were stored in a NoSQL 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 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 implemented for PISA 2022 was that it could be run without administrator rights to the local computer. This was a big improvement over earlier PISA cycles, thus significantly reducing greatly the amount of technical support needed within the schools.

## Data capture and scoring student responses

Student responses and other process data 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 sessions. At the end of a session, the results from that session were exported in a single password protected ZIP file. For the PISA tests from Session 1 (the cognitive PISA domains, including the optional financial literacy domain), 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.

In addition to these five standard variables, some more complex items 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 2022 was coding of student responses. For computer delivered tests, many of the item responses could be coded automatically. In PISA 2022, this included multiple-choice items, drag-and-drop items, numeric-response items, and complex responses to mathematics or science simulations.

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 conditional statements 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/economy. For PISA 2022, 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 units for PISA 2022, approximately 30% of the items required a response that needed to be coded by a human scorer or coder. On paper, this would be done directly on 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 example, for some items, a student would be required to 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 2022, the coding of these responses was done using the open-ended coding system (OECS). The OECS is a computer tool that was developed to support the coders in their work to code the responses according to the coding guides. All PISA 2022 open-ended responses collected with the computer-based platform were coded using the OECS.

The OECS works online so it required coders to have a reliable network connection. The OECS organizes responses according to the coding designs for each of the assessment domains. The system gives coders access to all the responses assigned to the coder. For each response, the coder will have access to part of the question for reference, the individual response to be coded, and the acceptable codes for each question. The coder selects the appropriate code and clicks on the "Record Code" button to save the selected code. It should be noted that this system was only used for response data from the computer-based assessment.

Also included on each page of the OECS were two checkboxes labelled "recoded" and "defer." The recoded box was used when the response had been recoded by another coder. The defer box was used when the coder was not sure what code to assign to the response. These deferred responses were typically reviewed and coded by the Lead Coder, or by the coder after consultation with the Lead Coder. When deferring a code for a response, coders were encouraged to enter comments into the box labelled "comment" to indicate the reason for deferring.

The OECS included the necessary features to support the monitoring of reliability. It organized all anchor, multiple and single coding of responses. According to a predetermined design, some responses were single coded – coded by one person only – while others will be multiple coded – coded by more than one coder. Anchor responses (in English) were used to assess reliability across countries. Since the OECS gives coders only those responses that are assigned to them, coders do not know whether they are single or multiple coding. Once coding was complete for each item, the data was integrated across coders and the OECS generated reliability reports that included multiple sections such as i) a summary, ii) item overview, iii) coders overview, iv) proportion agreement, v) coding category distributions, and vi) deferred and uncoded report.

# 22 International data products

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 from [www.iea.nl](http://www.iea.nl).

## Public-use files

The public-use files (PUF) contain response records from all participating countries/economies that are part of the approved PISA sample. Student-level files contain over 6000 variables that include responses to the background questionnaire and the cognitive assessments, as well as sampling weights, proficiency estimates and variables derived from responses to the background questionnaire. The student and teacher files contain over 1000 variables. The variables included in the PUF represent a common subset of the variables that were collected across all participating countries/economies and are available on the OECD website at <http://www.oecd.org/pisa/>.

### ***Variables excluded or suppressed for some or all countries***

The PUF include only a subset of the variables included in the individual country files. The PUF do not include any data collected using national adaptations and extensions. Rather, they include common data that were collected or derived across all countries. Additionally, variables were also excluded after 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 PUF are:

- direct, indirect, and operational identifiers for respondents;
- certain background questionnaire (BQ) or process variables such as free text entry responses and random numbers used by the SDS to determine routing;;
- all national adaptations and extensions in the BQ;
- original scale score values (theta) before standardisation to an international metric.

Countries were given the option of suppressing variables in the PUF. Suppression of variables was approved when data presented a risk to student, school, and/or teacher anonymity. Suppressed data are represented in the database by means of missing codes.

### ***Data files***

Data files are provided in both SAS and SPSS formats. The files include:

- Student questionnaire data file: This file includes ID variables, all student questionnaire response data, parent-questionnaire response data, student and parent background questionnaire scale and

derived variables, plausible values for the core domains (Reading, Math, and Science), and overall and replicate student weights.

- School questionnaire data file: The school questionnaire data file includes ID variables, school questionnaire response data, school questionnaire scale and derived variables, and an overall school weight.
- Teacher questionnaire data file: The teacher questionnaire data file includes ID variables, teacher questionnaire response data, and teacher questionnaire scale and derived variables, and overall and replicate teacher weights.
- Cognitive item data file<sup>1</sup>: The cognitive data file includes ID variables, raw and coded item responses, item log data for the computer-based assessment (e.g., total time and number of actions) for the core domains (Mathematics, Reading, Science).
- Creative Thinking cognitive data file: The cognitive data file includes ID variables, Creative Thinking raw and coded item responses, computer-based assessment (CBA) item log data (total time and number of actions); and Creative Thinking plausible values including the Maths, Reading, and Science plausible values that were created as part of the population model with the Creative Thinking cognitive data.
- Financial Literacy student questionnaire data file<sup>2</sup>: This file includes ID variables, all student questionnaire response data, parent-questionnaire response data, student and parent background questionnaire scale and derived variables, plausible values for the domains assessed (Financial Literacy, Reading, and Maths), and overall and replicate student weights for the optional financial literacy sample.
- Financial Literacy cognitive item data file<sup>1 2</sup>: The cognitive data file includes ID variables, raw and coded item responses, item log data for the computer-based assessment (e.g., total time and number of actions) for the domains assessed in the Financial Literacy sample (Financial Literacy, Maths, Reading).
- Questionnaire timing data file: The questionnaire timing data file includes CBA questionnaire log data (i.e., total time on a unit/screen).

The Creative Thinking datasets and Financial Literacy datasets are scheduled to be published in 2024.

### ***Variables used in sampling, weighting and merging***

The variable STRATUM is included to identify sampling strata. The variable is created as a concatenation of a three-letter country code and a two-digit original stratum identifier.

The variables W\_FSTUWT and W\_FSTURWT1 - W\_FSTURWT80 represent the full student sampling weight, and the 80 replicate weights used for estimation of sampling variance.

The variable SENWT is a normalised weight variable typically used for analyses of student performance across a group of countries where contributions from each of the countries in the analysis is desired to be equal regardless of their population or sample size. The senate weight adds to a constant of 5 000 across all cases within each country/economy in the file. This weight adds to 5000 within each country/economy only when there is no missing data for the variable of interest. The relative contribution of each country/economy is affected by the incidence 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 school ID number, making it unique across all countries.

## Codebooks for the PISA 2022 public-use data files

Included with the PISA 2022 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 categorical variables from the cognitive and background questionnaire variables, as well as those that have been derived and/or added during data processing. The codebooks are available from the OECD website (<https://www.oecd.org/pisa/data/>).

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 the range of valid values (minimum and maximum) for non-categorical numeric variables.

Codebooks for the main files are contained in separate worksheets within the file made available at the OECD website. Each worksheet corresponds to one of the eight public-use data files described above.

### ***Data compendia tables***

Using the PUF as the source data, the compendia are sets of summary tables that provide 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 to confirm that they are performing analyses on the PUF 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 with 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 items). 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.

Separate tables are provided with percentage and percentile data for continuous background variables across all questionnaires.

All statistics including in the compendia are calculated using weighted data and are presented with their corresponding standard error that take into account both the sampling and measurement uncertainty. The OECD average is created as the simple average of the 38 current OECD member countries.

## 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 the IEA's International Database Analyzer (IDB Analyzer). Each of these two software tools address 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 the PISA 2022 data, the PDX database contains data from previous cycles of PISA. The PDX is available on the OECD website (<https://pisadataexplorer.oecd.org/ide/idepisa/>) and provides access to a secure PISA database that is protected by the OECD firewalls and security mechanisms. The PDX allows the user to navigate, analyse, and produce report quality tables and graphics.

The database underlying the PDX is populated using the PUF to import more than 3.5 million unique student records across eight PISA cycles. About 8,700 variables across eight assessment cycles and over 100 countries, economies, 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 in the PUF but not the PDX relate to the individual cognitive item responses 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.

In the PISA Data Explorer, the International Average (OECD) includes all OECD member countries for which data are available for the corresponding subject and year (38 OECD Member countries as of PISA 2022). Depending on data availability, the countries contributing to this average might vary by cycle and subject.

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 variables selected, 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 Microsoft Edge, Firefox, Google Chrome, and Safari.

### *Import of trend data*

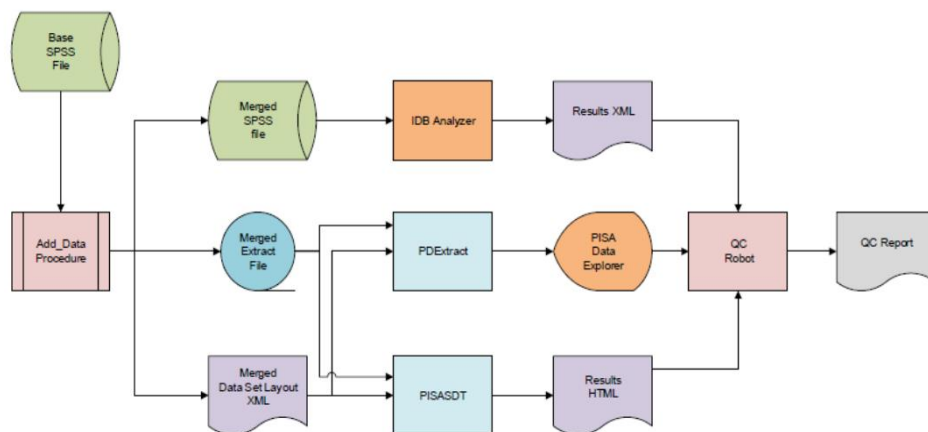
The PISA trend data from 2000 to 2012 were imported into the PDX directly from a database that had been established earlier by the United States Department of Education to develop and support a Data Explorer for PISA and other international studies. These data were taken from all PUF that were available 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 variables in subsequent PISA cycles (2015, 2018, and 2022) in order to ensure continuity and comparability with previous cycles.

## 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 22.1 below. This process was applied separately to the data from each country.

**Figure 22.1. 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 PDEExtract 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 has produced and is available upon request to the OECD Secretariat<sup>3</sup>. 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.

## IEA's International Database Analyzer

The IEA International Database Analyzer (IDB Analyzer) is an application developed by the International Association for the Evaluation of Educational Achievement (IEA) that can be used to analyse data from most major large-scale assessment surveys, including those conducted by the OECD, such as PISA. Originally designed for international large-scale assessments, it is also capable of working with national assessments such as the United States National Assessment of Educational Progress (NAEP).

The IDB Analyzer creates SPSS, SAS, or R syntax that can be used to perform analysis with these international databases. The syntax considers information from the sampling design in the computation of sampling variance and handles the multiple plausible value imputations. 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, SAS, or R macros, and users are only authorised to use the SPSS, SAS, and R 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:

- 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.
- Percentages only: Computes percentages by subgroups defined by the user.
- 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.
- 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.
- 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 as well as the computation of group mean and percent differences between groups taking into account sample dependency.
- 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.
- Percentiles: Computes the score points that separate a given proportion of the distribution of scores by subgroups defined by the grouping variable(s).

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, SAS, or R 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.

## Notes

---

<sup>1</sup> After the analysis phased completed, it was determined that 4 students in Iceland's grade-based sample were analysed along with Iceland's main sample data. As a result, the public use data for Iceland excludes these 4 students, yet they are still included in PISA 2022 technical report tables where Iceland data are referenced.

<sup>2</sup> For Financial Literacy, only a subset of participants for Canada and Belgium received the Financial Literacy assessment and it is not a nationally representative sample. Only the Belgium Flemish community as well as the Canadian provinces British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario and Prince Edward Island, participated in Financial Literacy for PISA 2022.

3. via email to [EDU.Pisa@oecd.org](mailto:EDU.Pisa@oecd.org)

## Annex A. Item Pool Classification tables

**Table A A.1. Item Pool Classification tables**


Table	Title
Table A.1	Math - Computer-based assessment
Table A.2	Math - New paper-based assessment
Table A.3	Math - Old paper-based assessment
Table A.4	Reading - Computer-based assessment
Table A.5.	Reading Fluency - Computer-based assessment
Table A. 6	Reading - New computer-based assessment
Table A.7	Reading Components - New paper-based assessment
Table A.8	Reading - Old paper-based assessment
Table A.9	Science - Computer-based assessment
Table A.10	Science - New paper-based assessment
Table A.11	Science - Old paper-based assessment

StatLink  <https://stat.link/5nw02t>

## Annex B. Contrast Coding Tables

Table A B.1. Contrast Coding Tables

Table B.1	Contrast Coding BQ Variables
-----------	------------------------------

StatLink  <https://stat.link/gqj28y>

# Annex C. Student and School Sample Size Tables

**Table A C.1. Student and School Sample Size Tables**

Table C.1	Main Sample Sizes by Country Domain
Table C.2	Financial Literacy Sample Sizes by Country and Domain

StatLink  <https://stat.link/nzriq>

# Annex D. National Household Possession Items Tables

Table A D.1. National Household Possession Items

	Variable Name			
	ST250Q06JA	ST250Q07JA	ST251Q08JA	ST251Q09JA
OECD Countries				
Australia	a home theatre	a pool or outdoor spa bath/jacuzzi	air conditioning unit	solar panels
Austria	swimming pool/pond	n/a	n/a	n/a
Belgium	a room where you can study quietly	n/a	antiques	n/a
Canada	gaming console (e.g. Nintendo Switch™, Xbox®, PlayStation®)	your own sports equipment	Smart home devices (e.g. smart thermostat, Google Home™, Amazon Echo Dot™)	Television or video subscription service (e.g. cable TV, Netflix®, Apple TV®)
Chile	printer	scanner	digital video camera	exercise machines that are working
Colombia	television	refrigerator	video game console	n/a
Costa Rica	your own TV	3D screen	cars	3D screens
Czech Republic	n/a	n/a	n/a	n/a
Denmark	your own game console	your own headphones	boat	n/a
Estonia	game console (PlayStation®, Xbox®)	your own table for studying	n/a	n/a
Finland	alarm system	garage or carport	n/a	n/a
France	a paying television program (e.g. Netflix, Canal Plus, OCS)	an action camera (e.g. GoPro)	n/a	n/a
Germany	a desk to study at	a quiet place to study	n/a	n/a
Greece	iPod and MP3 players	digital games (e.g. Playstation4 ®)	dishwasher	home alarm system
Hungary	air conditioner	dishwasher in kitchen	Video game console (e.g. Sony PlayStation™)	Digital camera (not built in a cell phone)
Iceland	n/a	n/a	n/a	n/a
Ireland	n/a	n/a	Subscription to TV or streaming services	n/a
Israel	4WD car	membership to the theater, gym, or swimming pool	n/a	n/a
Italy	printer	n/a	air conditioners	n/a
Japan	game console (e.g. PlayStation 4®, Nintendo Switch™)	passport	air conditioner	rooms for visitors
Korea	a desk to study at	massage chair	air conditioner	air cleaner
Latvia	bicycle	scooter	antique things	textile works
Lithuania	n/a	n/a	n/a	n/a
Mexico	a desk to study at	a quiet place to study	n/a	n/a
Netherlands	your own personal computer or laptop	your own tablet (e.g. iPad, Samsung Galaxy)	a subscription to a newspaper	an electric car

	Variable Name			
	ST250Q06JA	ST250Q07JA	ST251Q08JA	ST251Q09JA
New Zealand	your own snowboard or skis	your own musical instrument (e.g. guitar, keyboard)	heat pumps	Large outdoor recreation items (e.g. tent, boat, mountain bike, surfboard)
Norway	a good place to do school work	n/a	electric bicycles	n/a
Poland	n/a	n/a	washer	n/a
Portugal	cable TV or satellite TV	n/a	poetry books	n/a
Slovak Republic	n/a	n/a	n/a	n/a
Slovenia	multifunction printer	sports equipment (e.g. skis, bike, tennis racket, etc.)	external hard disk	sauna
Spain	media services for TV series and films (HBO, Netflix)	pay TV (Movistar+, Orange TV)	dishwashers	parking places
Sweden	n/a	n/a	home cinema	n/a
Switzerland	n/a	n/a	dishwashers	mowing machines
Türkiye	TV subscriptions with payment (e.g. Digiturk, Tivibu and Teledunya)	helper for houseworks (part time or full time)	Air conditioning type heating-cooling system	LCD, LED TV or Plasma TV
United Kingdom (Excl. Scotland)	a games console such as a PlayStation 4® or Nintendo Wii®	a smart speaker such as Amazon Echo or Google Home	a study desk or table for your use	computers (e.g. desktop, laptop or tablet)
United Kingdom (Scotland)	your own bicycle	your own smartwatch	spaces to park cars	outdoor spaces attached to your home (e.g. garden)
United States	n/a	n/a	n/a	n/a
Partner Countries/Economies				
Albania	n/a	n/a	n/a	n/a
Argentina	a quiet place to study	n/a	n/a	n/a
Baku (Azerbaijan)	washing machine	n/a	n/a	n/a
Brazil	cable TV	your own desk to study at	game console with internet access	refrigerator
Brunei Darussalam	bedroom with an air conditioner	Video or online games (e.g. used with game consoles such as a PlayStation 4®)	rooms with marble floor	surveillance camera or CCTV
Bulgaria	digital camera	air conditioning	n/a	n/a
Cambodia	books to help with your school work	a dictionary	refrigerator	smart televisions ( Internet connectivity)
Croatia	n/a	n/a	dishwasher	air conditioner
Cyprus	a swimming pool	a home security alarm system	Cable or Satellite TV (e.g. Cablenet, Cytavision, Nova, PrimeTel)	Game Platforms (e.g. Playstation, Nintendo Switch/Wii)
Dominican Republic	wrist watch	your own car	air conditioner	smart TV
El Salvador	typewriters	microwave	trees	pets
Georgia	video games	n/a	n/a	n/a
Guatemala	stereo	blender	books to help with your school work	n/a
Hong Kong (China)	storeroom	Newspaper or Educational Magazine (e.g. National Geographic Magazine)	television	air conditioning unit
Indonesia	your own Personal Computer/ Desktop / Laptop	your own tablet/iPad	refrigerator	oven
Jamaica	cable TV	portable Wi-fi	TV	car
Jordan	iWatch	digital books	antiques	office rooms
Kazakhstan	bicycle	digital photo camera	bicycle	digital photo camera




	Variable Name			
	ST250Q06JA	ST250Q07JA	ST251Q08JA	ST251Q09JA
Kosovo	n/a	n/a	n/a	n/a
Macao (China)	safe box	electric massage chair	air purifier	hi-fi audio set
Malaysia	printer	refrigerator	pressure cooker	air conditioning unit
Malta	n/a	n/a	smart TV with internet access	n/a
Mongolia	n/a	n/a	silver bowl	carpet
Montenegro	n/a	n/a	n/a	n/a
Morocco	swimming pool	electric water heater tank	smart TV	fishing boat
North Macedonia	n/a	n/a	n/a	n/a
Palestinian Authority	iWatch	digital books	antiques	office rooms
Panama	all in one printer	simulation tools	cable or satellite TV	internet TV
Paraguay	n/a	n/a	n/a	n/a
Peru	Playstation, Nintendo, Wii	bike	refrigerator	stereo
Philippines	means of transportation (e.g. motorcycle, tricycle, jeepney, car, etc.)	air conditioning unit	Video game console (PlayStation, Xbox, etc.).	smart TV
Qatar	office	cinema	digital video camera	video game console
Republic of Moldova	n/a	n/a	n/a	n/a
Romania	air conditioning	cable/satellite TV	n/a	n/a
Saudi Arabia	gaming system	n/a	n/a	n/a
Serbia	n/a	n/a	LED/LCD/Plasma TV	digital camera
Singapore	air conditioner	domestic helper (e.g. full/part-time maid)	n/a	n/a
Chinese Taipei	n/a	n/a	n/a	n/a
Thailand	smart television	air purifier	air conditioner	refrigerator
Ukraine	your own desk to do hometasks	reference books	family heirlooms	smart TV
United Arab Emirates	cinema	electronic games with internet access	luxury cars (e.g. Bentley, Rolls-Royce)	domestic Workers (e.g. housemaid, Drivers, etc.)
Uruguay	a desk to study	n/a	smart TV	n/a
Uzbekistan	TV with an access for international channels	bookshelf	bicycle	refrigerator
Viet Nam	a dictionary	a set of chair and table for learning	air conditioner	n/a

## Annex E. Final Distribution of RMSD Values Across Groups for Each Scale Tables

Table A E.1. Final Distribution of RMSD Values Across Groups for Each Scale

Table E.1	RMSD values for BQ scales
-----------	---------------------------

*StatLink*  <https://stat.link/unoab8>

# Annex F. Common and Unique Item Parameters in Each Domain, by Country and Language Tables

**Table A F.1. Common and Unique Item Parameters in Each Domain, by Country and Language**

Table F.1	Consolidated common unique item parameters
Table F.2	Summary
Table F.3	Unique item parameters

StatLink  <https://stat.link/efz3n0>

# Annex G. Equated P Tables

Table A G.1. Equated P Tables


Table G.1	Guidelines
Table G.2	Equated P+
Table G.3	Equated P+ (Standard error)
Table G.4	Country economy code 3-character

StatLink  <https://stat.link/5dqy71>

# Annex H. Testing Periods Tables

**Table A H.1. Testing Periods**

Table H.1	Testing periods worksheet
-----------	---------------------------

StatLink  <https://stat.link/l2x4p5>

# Annex I. PISA 2022 Technical Standards and guidelines

## Purpose of document

The purpose of this document is to list the set of standards upon which the PISA 2022 data collection activities will be based, as was the case for previous PISA. 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 that were translated and/or adapted as appropriate. Comparable samples of each country's 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 are questionable. 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 should be selected, and test materials and tasks etc. be developed 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 be essential, but not sufficient, for the results to 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 harmonised at a later stage in the project. Therefore, the standards specify one clear-cut path along which data collection, coding 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, and the quality assurance and quality management plan to demonstrate that the standard has been met. While the terms quality assurance and quality control are sometimes used interchangeably, they relate to different aspects

of quality. Quality assurance is most often associated with the processes and procedures that are put in place to make sure the survey is likely to meet its intended goals. Quality control, on the other hand, relates to the set of judgements that are made with regard to the suitability of the quality assurance procedures and the suitability of the survey results in terms of their intended uses or applications.

Where standards have been fully met and data quality of the final databases judged as appropriate, the international contractors will recommend to the OECD Secretariat that the data be included in the PISA 2022 database. Where standards have not been fully met or data quality has been questioned, an adjudication process will determine the extent to which the quality and international comparability of the data have been affected or whether additional analysis or evidence are necessary. The result of data adjudication will determine whether the data will be recommended for inclusion in the PISA 2022 dataset.

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 testing, printing 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 or data quality are identified, the International Survey Director initiates communication with the National Centre as soon as possible to give advice on resolving problems.

The PISA standards serve as benchmarks of best practice. As such, the standards are designed to assist National Centres and the 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 contractors are defined in the PISA operations manuals.

As specified in the contracts for the implementation of the eighth cycle of the OECD Programme for International Student Assessment, the international contractors take responsibility for developing and implementing procedures for assuring data quality. Therefore, the international contractors mediate, and monitor 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. The international contractors must communicate timelines and tasks well in advance to National Centres and give reasonable deadlines for National Centres to respond to tasks.

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 data quality and its assurance.
- These standards have been reviewed by the Technical Advisory Group, and their comments and suggestions have been taken into careful consideration in finalising the standards.
- 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 2022 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 control 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

### ***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. In PISA 2022, a teacher questionnaire will be offered as an international option. The response-rate standard for teachers specified in this section applies only to countries that participate in this international option, and will ensure that the analysis and reporting goals for this option can be met.

The procedures for drawing the samples used in the study are 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 results are comparable, reliable and valid. To reach these goals the sampling procedures must follow established scientific principles for drawing samples from finite populations.



**Standard 1.1** The PISA Desired Target Population is agreed upon through negotiation between the National Project Manager and the international contractors within the constraints imposed by the definition of the PISA Target Population. The Target Population for PISA starts with students attending all educational institutions located within the country, and in grade 7 or higher. The “standard” PISA target population is further refined to its age basis: students between 15 years and 3 (completed) months and 16 years and 2 (completed) months at the beginning of the testing period.

**Standard 1.2** Unless *otherwise agreed upon* only *PISA-eligible students* participate in the test.

**Standard 1.3** Unless otherwise agreed upon, the testing period:

- is no longer than eight consecutive weeks in duration for computer-based testing participants,
- is no longer than six consecutive weeks in duration for paper-based testing participants,
- does not coincide with the first six weeks of the academic year, and

begins exactly three years from the beginning of the testing period in the previous PISA cycle

**Standard 1.4** Schools are sampled using agreed upon, established and professionally recognised principles of scientific sampling.

**Standard 1.5** Student lists should not be collected more than 8 weeks prior to the start of data collection, unless otherwise agreed upon.

**Standard 1.6** Students are sampled using *agreed upon*, established and professionally recognised principles of scientific sampling and in a way that represents the full population of *PISA-Eligible students*.

**Standard 1.7** The PISA Defined Target Population covers 95% or more of the PISA Desired Target Population. That is, school-level exclusions and within-school exclusions combined do not exceed 5%.

**Standard 1.8** The student sample size for the **computer-based mode** is a minimum of 6 300 assessed students, and 2 100 for additional adjudicated entities, or the entire PISA Defined Target Population where the PISA Defined Target Population is below 6 300 and 2 100 respectively. The student sample size of assessed students for the **paper-based mode** is a minimum of 5 250.

**Standard 1.9** The school sample size needs to result in a minimum of 150 participating schools, and 50 participating schools for additional adjudicated entities, or all schools that have students in the PISA Defined Target Population where the number of schools with students in the PISA Defined Target Population is below 150 and 50 respectively. Countries not having at least 150 schools, but which have more students than the required minimum student sample size, can be permitted, if agreed upon, to take a smaller sample of schools while still ensuring enough sampled PISA students overall.

**Standard 1.10** The minimum acceptable sample size in each school is 25 students per school (all students in the case of school with fewer than 25 eligible students enrolled).

**Standard 1.11** The final weighted school response rate is at least 85% of sampled eligible and non-excluded schools. If a response rate is below 85% then an acceptable response rate can still be achieved through *agreed upon* use of replacement schools.

**Standard 1.12** The final weighted student response rate is at least 80% of all sampled students across responding schools.

**Standard 1.13** The final weighted teacher response rate is at least 75% of all sampled teachers across responding schools.

**Standard 1.14** The final weighted sampling unit response rate for any optional cognitive assessment is at least 80% of all sampled students across responding schools.

**Standard 1.15** Analyses based on questionnaire data that do not link to a weighted 75% of the target population shall be flagged or replaced by a missing code in OECD reports.

**Standard 1.16** Unless *otherwise agreed upon*, the international contractors will draw the school sample for the Main Survey.

**Standard 1.17** Unless *otherwise agreed upon*, the National Centre will use the sampling contractor's software to draw the student sample, using the list of eligible students provided for each school.

Note 1.1 Standards 1.1 through 1.17 apply to the Main Survey but not the Field Trial.

Note 1.2 Data from schools where the (unweighted) student response rate is greater than 33% will be included in the PISA dataset and the school counted as a respondent. Otherwise, the school will be a non-respondent, and no student, school or teacher data will be retained.

Note 1.3 A PISA-eligible student recorded in the database as not doing the minimum required number of questions of the main cognitive part of the PISA assessment will be counted as a nonparticipant.

Note 1.4 Acceptable response rates obtained through the use of replacement schools are described in detail in the School Sampling Preparation Manual.

Note 1.5 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 whose students 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 1.6 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 from each sampled school: The recommended target cluster size is 42 and 25 is the minimum. In determining the target cluster size for a given country, or stratum within a country, it is necessary to ensure that the minimum sample size requirements for both schools and students will be met.

Note 1.7 Any exceptional costs associated with verifying a school sample taken by the National Centre, or a student sample selected other than by using the sampling contractor's software will be borne by the National Centre.

Note 1.8 Agreement with the international contractor on 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 the sampling contractor's software 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.

Note 1.9 Teachers recorded in the database as completing at least one valid response will be counted as respondents.

## Quality assurance

- Sampling procedures as specified in the PISA operations manuals
- School sample drawn by the international contractors (or if drawn by the National Centre, then verified by the international contractors)
- Student sample drawn through the sampling contractor's software (or if drawn by other means, then verified by the international contractors)
- Sampling forms submitted to the international contractors
- Main Survey Review Form

## 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.

**Standard 2.1** 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 (Mathematics) of the test.

If the language of instruction in the major domain is not well defined across the set of sampled students, then, if *agreed upon*, 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 *adjudicated entity*.

If the language of instruction differs across domains, then, if *agreed upon*, 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.

In all cases the choice of test language(s) in the assessment instruments is made prior to the administration of the test.

### **Field Trial participation**

**Rationale:** The Field Trial gives countries the opportunity to try out the logistics of their test procedures and allows the contractors to make detailed analyses of the items so that only suitable ones are included in the Main Survey.

**Standard 3.1** PISA participants participating in the PISA 2021 Main Survey will have successfully implemented the Field Trial. Unless otherwise agreed upon:

A Field Trial should occur in an assessment language if that language group represents more than 5% of the target population.

For the largest language group among the target population, the Field Trial student sample should be a minimum of 200 students per item.

For all other assessment languages that apply to at least 5% of the target population, the Field Trial student sample should be a minimum of 100 students per item.

For additional adjudicated entities, where the assessment language applies to at least 5% of the target population in the entity, the Field Trial student sample should be a minimum of 100 students per item.

**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 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 3.3** 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.

### ***Adaptation of tests, questionnaires and school-level materials***

**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 competencies 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 and other survey materials have to be as equivalent as possible. In fact, it is of utmost importance to provide equivalent information to 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 equivalent information and training in all participating countries.

**Standard 4.1** The majority of test items used in previous cycles will be administered unchanged from their previous administration, unless amendments have been made to source versions, or outright errors have been identified in the national versions.

**Standard 4.2** All assessment instruments are equivalent to the source versions. Agreed upon adaptations to the local context are made if needed.

**Standard 4.3** National versions of questionnaire items used in previous cycles will be administered unchanged from their previous administration, unless amendments have been made to source versions, outright errors have been identified in the national versions, or a change in the national context calls for an adjustment.

**Standard 4.4** The questionnaire instruments are equivalent to the source versions. Agreed upon adaptations to the local context are made if as needed.

**Standard 4.5** School-level materials are equivalent to the source versions. Agreed upon adaptations to the local context are made as 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 10% of the target population.

#### Quality assurance

- Agreed upon adaptation to school-level materials using methods specified by the international contractors
- Questionnaire Adaptation Spreadsheet (QAS)
- Test Adaptation Spreadsheets (TAS, for paper and computer instruments) or other agreed upon monitoring tool in which adaptations to assessment units, orientation and help files and coding guides are documented. For languages that are the languages of instruction for 10% or more of the target population, 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.
- For languages that are the languages of instruction for 10% or more of the target population: Verifier Reports (verification statistics generated by the monitoring tool, in combination with a short qualitative report)
- Field Trial and Main Survey Review Forms.
- Item and scale statistics generated by the international contractors (assessment materials and questionnaires).

### ***Translation of assessment instruments, questionnaires and school-level materials***

**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 competencies will be assessed, and not variations of information caused by differences in the translation of materials.

**Standard 5.1** The following documents are translated into the assessment language in order to be linguistically equivalent to the international source versions.

- All administered assessment instruments
- All administered questionnaires
- The Test Administrator script from the Test Administrator (or School Associate) Manual
- The Coding Guides (unless otherwise agreed upon)

**Standard 5.2** Unless otherwise agreed upon, school-level materials are translated/ adapted into the assessment language to make them functionally equivalent to the international source versions.

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.

#### Quality assurance

- Agreed upon Translation Plan, developed in accordance with the specifications in the PISA operations manuals, that requires double translation by independent translators followed by reconciliation for any newly translated questionnaires and cognitive instruments; and a thoroughly documented adaptation process for any materials adapted from one of the source versions, from a common reference version, or from verified materials borrowed from another country.
- Agreed Upon Questionnaire Adaptation Spreadsheet (QAS)
- Test Adaptation Spreadsheets (TAS) or other agreed upon monitoring tool in which adaptations to assessment units, orientation and help files 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 (verification statistics generated by the monitoring tool, in combination with a short qualitative report)
- Submitted final materials as used in the study
- Field Trial and Main Survey Review Forms
- Item and scale statistics generated by the international contractors (assessment materials and questionnaires)

### ***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.

**Standard 6.1** 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.

**Standard 6.2** 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. Testing results must be submitted to the international contractors so that any errors can be promptly resolved.

#### Quality assurance

- Detailed testing plans
- Review of testing results

### **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** In countries that administer the computer-based version of PISA, the 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 Forms

### **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 blank papers and calculators, the instructions given prior to testing, the rules for excluding students from the assessment, etc. A list of these and other relevant test conditions is given in the school-level materials. 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

instruments (tests and questionnaires) 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.

**Standard 8.1** All test sessions follow international procedures as specified in the PISA school-level materials, particularly the procedures that relate to:

- test session timing,
- maintaining test conditions,
- responding to students' questions,
- student tracking, and
- assigning assessment materials.

**Standard 8.2** 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, a relative, or a personal acquaintance of any student in the assessment sessions he or she will administer for PISA.

**Standard 8.3** National Centres must not offer rewards or incentives that are related to student achievement in the PISA test to students, teachers, or schools.

Note 8.1 Test Administrators should preferably not be school staff.

Note 8.2 This does not apply to incentives or rewards intended to improve participation, and that are unrelated to student achievement in the PISA test.

#### Quality assurance

- Session Report Forms
- PISA Quality Monitors feedback and Data Collection Forms (only for Main Survey)
- Field Trial and Main Survey Review Forms

#### ***Training support***

**Rationale:** NPMs or their designees shall participate in a train-the-trainer session conducted by qualified contractor staff. This facilitates 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.

**Standard 9.1** 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.

**Standard 9.2** NPMs or designees shall use the comprehensive training materials and approach developed by the contractors and provided on the PISA Portal to train PISA test administrators.

**Standard 9.3** 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 make results less comparable.

**Standard 9.4** In-person and/or web based test administrator trainings should be conducted by the NPMs or designees, unless a suitable alternative is *agreed upon*.

**Standard 9.5** PQMs need to successfully complete self-training materials, attend webinars to review and enhance the self-training, and attend the *test administrator* training, *unless otherwise agreed upon*.

#### Quality assurance

- Participation in trainer training sessions in standardised procedures by qualified contractor staff
- National Centre Quality Monitoring
- Field Trial and Main Survey Review Forms
- Standard training of PQMs
- Review of Test Administrator Training Observation Forms

#### ***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 components 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, unless otherwise agreed upon.

#### ***Security of the material and test preparation***

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. As high levels of student and school participation in PISA are very important, it is appropriate for national centres to prepare communication materials for participants with the intent to raise awareness, to set out what is involved in participating in PISA and to encourage participation in the survey. These materials may include general information about the survey, what students and schools might expect on the test day, as well as an OECD set of released test materials prepared for this purpose. The use of sample test items in informational materials could also serve to prepare students for the format of the PISA test in order to reduce potential test anxiety and help the students focus on the subject-matter content when taking the test.



**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 materials,
- 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.

**Standard 11.2** Participating schools, students and/or teachers should only receive general information about the test prior to the test session, rather than formal content-specific training. In particular, it is inappropriate to offer formal training sessions to participating students, in order to cover skills or knowledge from PISA test items, with the intention to raise PISA scores.

Note 11.1: It is unnecessary to train students for interacting with the student interface, with different item types or response formats prior to the testing session. All PISA test materials and procedures are accompanied by detailed instructions as well as by orientation modules at the beginning of each test session to ensure that participants are familiarised with the interface and with all the question formats that they will encounter.

Note 11.2: "Formal training sessions" refers to training that relies on standardised instructional material and involves feedback provided by an instructor, machine, or other training participants. Formal training sessions may include (but are not limited to) lectures, practice tests, drills or online instruction modules.

Note 11.3. The general information about the survey shared with participants may include information about the length of the test, the general scoring principles applied to missing and incorrect answers, data protection and confidentiality of results. It may include an OECD set of released test materials prepared for this purpose but should not assemble sample items in PISA-like test forms with the intent to teach or prepare students for participation in PISA.

#### Quality assurance

- Security arrangements as specified in the PISA operations manuals or agreed upon variation
- National Centre Quality Monitoring
- PISA Quality Monitor feedback and Data Collection Forms (only for Main Survey)
- Field Trial and Main Survey Review Forms

#### **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 for the Main Survey.

**Standard 12.1** PISA Main Survey test administration is monitored using site visits by trained independent quality monitors.

**Standard 12.2** Fifteen site visits to observe test administration sessions are conducted in each PISA participating country/economy, and five site visits in each adjudicated region.

**Standard 12.3** Test administration sessions that are the subject of a site visit are selected by the international contractors to be representative of a variety of schools in a country/economy.

Note 12.1 A failure to meet the Quality Monitoring standards in the Main Survey could lead to a significant lack of reliable and valid quality assurance information.

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 contractors the assistance required to implement the site visits effectively. This includes nominating sufficient qualified individuals to ensure that the required number of schools is observed. It also includes timely communication of school contact information and test dates.

#### Quality assurance

- The process of selecting the PISA Quality Monitor nominees .
- PISA Quality Monitor feedback and Data Collection Forms (only for Main Survey)

#### ***Assembling and printing paper-based materials***

Rationale: Variations in assembly and 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 participating in paper-based components in their first cycle of participation.

**Standard 13.1** All paper-based student assessment material will be centrally assembled by the international contractors and must be printed using the final print-ready file and *agreed upon* paper and print quality. New countries/entities must submit a printed copy of all Field Trial instruments (booklets and questionnaires) for approval of the printing quality for the Main Survey. The same printing standard must be used for both the Field Trial and the Main Survey.

**Standard 13.2** The cover page of all national PISA test paper-based materials used for students and schools must contain all titles and approved logos in a standard format provided in the international version.

**Standard 13.3** The layout and pagination of all test paper-based material is the same as in the *source versions, unless otherwise agreed upon*.

**Standard 13.4** The layout and formatting of the paper-based questionnaire material is equivalent to the source versions, with the exception of changes made necessary by national adaptations.

Note 13.1 The cover page of all PISA PBA 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

- Agreement that quality will be similar to Field Trial versions
- For new countries/economies, materials submitted to the international contractors, as described in Standard 13.1 above.
- Field Trial and Main Survey Review Forms

### **Response coding**

**Rationale:** To ensure the comparability of the data, the responses from all test participants in all participating countries have to be coded following approved coding designs that are presented to both the Field Trial and the Main Survey. 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 guides is implemented according to instructions from the international contractors' 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 must be implemented as specified in the *PISA operations manuals* (see Note 14.1). These procedures are implemented in the coding software that countries will be required to use.

**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 number of Coder Training session participants will 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 Forms

### **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 contractors. 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 contractors 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, questionnaires 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

### ***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 various resources provided by the contractors.

**Standard 16.1** The international contractors ensure that qualified staff are available to respond in English 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 in English by the international contractors during all stages of the project. The qualified staff:

- Are authorised to respond to queries,
- Are able to communicate in English,
- Acknowledge receipt of queries within one working day,
- Respond to queries from the 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 Tasks section of the PISA Portal.

### ***Notification of international and national options***

**Rationale:** Given the tight timelines, the deadlines given in the following two standards will enable the international contractors to progress with work on time.

**Standard 17.1** National options are agreed *upon* with the international contractors before 1 December in the year preceding the Field Trial and confirmed before 1 November in the year preceding the Main Survey.

**Standard 17.2** The National Centre notifies the OECD Secretariat of its intention to participate in specific international options three months prior to the start of the translation period. International options can only be dropped between the Field Trial and the Main Survey, not added.

### **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 new countries/entities using the paper-based instruments (where this is required, see Standard 13.1),
- results from the national checking of adapted computer-based assessment materials and questionnaires,
- adaptations to school-level materials,
- sampling forms (see Standard 1),
- demographic tables,
- completed Field Trial and Main Survey Review Forms,
- documents related to PISA Quality Monitors: nomination information, Test Administrator training schedules, translated school-level materials, school contact information, test dates, and
- 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 school-level materials that are required to be functionally equivalent to the source as specified in Standard 5.2, need to be *agreed upon*.

#### Quality assurance

- Agreed upon Translation Plan
- International contractors' records from communications, forms, or documents
- Assessment materials submitted for linguistic verification with corresponding adaptation spreadsheets filled in by the National Centre

### **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 eight 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 operations 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 communicate in English,
- is able to respond to international contractor queries within three working days, and
- 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 contractors on or before the first day of testing. The submission must include the: electronic PDF and/or Word versions of all 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 and delivered as agreed upon a pre-specified timeline that varies based on their data submission. The national database will contain the complete set of responses from the students, school principals and surveyed participants (parents, teachers) 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 international contractors 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 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, school-level materials, and coding guides) in any language and format.

## Quality assurance

- International contractors' records of communications, forms, or documents

## 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, school-level materials and coding guides, including all national versions. For documents that are finalised by countries, they are required to upload the latest version to the PISA Portal.

**Standard 20.2** The National Project Manager must submit one copy of each of the following adapted and translated Main Survey materials to the international contractors:

- electronic versions (Word and/or PDF) of all administered Test Instruments, including international and *national options*
- electronic versions (Word and/or PDF) of all administered Questionnaires, including international and *national options* (paper-based countries only);
- electronic versions of the school-level materials; and
- electronic versions of the Coding Guides.

**Standard 20.3** Unless otherwise requested, National Centres will retain (1) all Field Trial materials until the beginning of the Main Survey, and (2) all Main Survey materials until the end of the calendar year, two years after the year when the Main Survey is conducted, (i.e. when the last international reports containing the results of the Main Survey will have been published). Materials to be archived include:

- all respondents' paper-based test booklets and questionnaires (PBA countries or whenever paper-based materials are used in CBA countries)
- all respondents' SDS result files and all associated data obtained from USB drives or other delivery mode (CBA countries)
- all sampling forms,
- all respondent lists,
- all tracking instruments, and
- all data submitted to the international contractors.

After completion of a survey, the National Centre will transfer final versions of all national materials to the international contractors who will compile the national archives from all participants and transfer them to OECD after completion of the Main Survey.

Note 20.1. Archiving applies to all materials from the Field Trial or Main Survey, including student, school, parent and teacher materials, as applicable.

Note 20.2. Should national legislation or other circumstances require that the Field Trial or Main Survey materials be deleted/erased before the timeline in Standard 20.3, countries must nevertheless retain these records, at a minimum, until the publication of the PISA dataset (and publication of the related international reports).

Note 20.3. It is recommended to retain the original USB drives (if used) and all paper-based booklets and questionnaires for all respondents until certified data has been released to the National Centres. Original USB drives are not required for long-term archiving purposes as long as there are copies of SDS result files for all respondents.

Note 20.4. Sampling forms for each sampling task for the Field Trial and Main Survey data collections must be retained for the periods outlined in Standard 20.3.

Note 20.5. "Respondent lists" refers to the student list (and teacher list if applicable) used for within-school sampling, and must be retained until, at a minimum, the period set out in Note 20.2.

Note 20.6. "Tracking instruments" refers to the Student Tracking Form (and Teacher Tracking Form if applicable) completed in each school, and must be retained until, at a minimum, the period set out in Note 20.2.

## ***Data protection and the processing of personal data***

**Rationale:** The OECD is committed to protecting the personal data it processes, in accordance with its Personal Data Protection Rules. The OECD, countries and contractors must protect the personal data of participants collected during PISA, ensuring that all data is stored and processed in a secure and standardised manner. This standard aims to ensure that National Centres process personal data securely, that participants are provided with clear information on data protection in PISA and the data rights of participants to access, rectify or erase their data are facilitated by countries and contractors.

**Standard 21.1** Each National Centre must make data protection information available to all participants, that at least includes:

- Contact details of the National Centre
- Contact details of the OECD's Data Protection Officer and Data Protection Commissioner
- The purpose of the processing of the data
- Recipients of the data, including any international organisations or third party (this includes the PISA contractors and any national contractors)
- The storage and retention period of data
- The existence of the rights of data subjects, including the timeline for facilitating these requests.

**Standard 21.2** Each National Centre will process the additional information related to a participant/data subject (e.g. link files with records of student names) securely and separately from the datasets during data processing and archiving (i.e. the data collected during the assessment will be categorised and treated as pseudonymised).

**Standard 21.3** National Centres must inform schools or other holders of PISA forms that include student and/or teacher names, to delete, confidentially shred or return these materials to the National Centre. Materials to be deleted/shredded/returned include:

- All tracking instruments
- All respondent lists.

**Standard 21.4** Each National Centre must facilitate requests from participants to exercise their data rights.

- Data access requests will be possible using the raw data from the assessment. No scaled data will be provided in breach of the PISA data embargo.
- Data erasure requests will be possible for a limited period before submission to the Contractors. This is to be decided by each National Centre, with two options, up to the submission of ST12 or to upload of student data files to the OECS.



- Each National Centre will retain and update a log of completed data requests for data erasure, to facilitate quality control processes. This information must be submitted to the PISA contractors in a timely manner to comply with the requests and for the purpose of data management and sampling processes.

Note 21.1. It is best practice to make data protection information available to participants at the time of the data collection.

Note 21.2. National Centres may communicate data protection information to participants in the most effective way for their national context. The information may be provided in several ways, e.g. a video, an information sheet, a data protection notice, on a National Centre website or a link to the OECD's PISA 2022 data protection notice.

Note 21.3. "Tracking instruments" refers to the Student Tracking Form (and Teacher Tracking Form if applicable) completed in each school. "Respondent lists" refers to a list of students (or list of teachers if applicable) used during the implementation of PISA.

Note 21.4. The records of student/teacher names in link files or on PISA forms are permitted but must be stored separately and securely to the data collected during the assessment.

Note 21.5. The data collected from students as part of PISA is categorised as pseudonymised, as identifying characteristics in the data have been replaced with a number or value that does not allow the data subject to be directly identified. This also pertains to data from parents or teachers, if applicable. Pseudonymisation means that the data collected remains personal data, but can no longer be assigned to a natural person without additional information (e.g. record of a student name and the PISA student ID number).

Note 21.6. After the archiving period for the Field Trial and Main Survey materials, National Centres may choose and are encouraged to anonymise the data by breaking the link between the name of the student and the data from the cognitive and questionnaire sessions. Anonymisation of the data requires deleting and/or confidentially shredding all files and records that connect the PISA student ID number to identifying information (name, date of birth, national student ID, etc.). Once anonymisation is complete and all records of a participant's name are removed, it will be no longer possible to facilitate access and erasure requests.

Note 21.7. The timeline of disposal/return of PISA forms retained by schools is to be decided by the National Centre and communicated to schools. This should be before the end of the archiving period set out in Standard 20.3.

#### Quality assurance:

- Adherence to this standard is a National Centre responsibility.
- Retain a copy of national data protection information for PISA 2022 made available to respondents.
- Agreement that additional information related to data subject (e.g. linking information) will be stored separately and securely from datasets.
- Agreement to set and follow-up on the timeline and procedures for the deletion and/or shredding of data in PISA forms held by other parties involved in the implementation of PISA.
- Retain a record of completed data access or erasure requests and submit requests to international contractors in a timely manner.

## National involvement standards

### **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 22.1** National Centres develop appropriate mechanisms in order to promote participation, effective implementation, and dissemination of results amongst all relevant national stakeholders.

**Standard 22.2** 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 22.1 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
- List of committees and groups of stakeholders
- Membership records of representative groups and/or committees
- Meeting records of representative groups and/or committees

**Meeting attendance**

**Rationale:** Attendance at National Project Managers and training meetings is required as these represent a key component of participating in PISA. Important information is shared and discussed and training in data management, sampling, computer systems, and coding is conducted at these international meetings. These also allow for individual consultation and communication with the international contractors, which is often very helpful.

**Standard 23.1** Representatives from each National Centre are required to attend all PISA international meetings including National Project Manager meetings, coder training, and any separate within-school sampling training, and data management training, as necessary. Up to 6 international meetings are planned per cycle.

**Standard 23.2** Representatives from each National Centre who attend international meetings must be able to work and communicate in English.

Note 23.1 The length of these meetings vary from 3 to 5 days.

Note 23.2 Based on the meeting type and hotel arrangements, the OECD Secretariat may, on the request of the international contractors, set a limit to the number of representatives per country that can attend NPM meetings. Countries/economies with separate participating entities will have the possibility to send teams from all entities.

## Quality assurance

- Meeting attendance records

**Definitions**

**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. A PISA participant may manage more than one adjudicated entity.

**Agreed procedures** – procedures that are specified in the PISA operations manuals, or variations that are mutually 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 mutually agreed upon between the National Project Manager and the international contractors.

**Agreed upon** – variations that are mutually agreed upon between the National Project Manager and the international contractors

**Anonymisation** - personal data is rendered anonymous, by irreversibly removing the link between each respondent's personal identifier (e.g. respondent name) and the data in the PISA dataset. This is achieved by deleting and erasing all additional information sources containing the link, so respondents in the PISA dataset cannot be personally identified. This applies to students, parents or teachers, if these options are administered, and anonymisation can be pursued after the archiving period in Standard 20.3.

**Centrally produced reference documents** – documents provided in English (and, for some documents, French and/or Spanish) by the international contractors according to contractual specifications.

**Common reference version** – a language version of assessment instruments that is used by countries sharing that language as a starting point to produce their respective national versions.

**International options** – optional additional international instruments or procedures sponsored by the OECD and fully supported by the international contractors.

**National Centre quality monitoring** – the procedures by which the international contractors monitor the quality of all aspects of the implementation of the survey by a National Centre.

**National option** – a national option occurs if:

- 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

- b) 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.

OR

- c) National Centre administers any PISA international option only in some, not all, jurisdictions. The country will in this case sign up for the international option with the OECD, as if it was administered in the entire jurisdiction, and the additional work involved with administering the international option to part of the jurisdiction only is considered a national option.

**PISA-eligible students** – students who are in the PISA target population. Also see PISA Target Population.

**PISA National Project Manager (NPM)** – The NPM is responsible for overseeing all national tasks related to the development and implementation of PISA throughout the entire cycle. The NPM is responsible for ensuring that tasks are carried out on schedule and in accordance with the specified international standards.

**PISA Operations Manuals** – all manuals provided by the international contractors. 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 or persons, usually the National Project Manager, who is/are responsible for administering PISA in one or more adjudicated entities. The National Project Manager(s) 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 Portal** – the PISA 2022 project website can be accessed through the following address: <http://pisa.ets.org/portal>.

**PISA Quality Monitor (PQM)** – a person nominated by the National Project Manager and employed by the international contractors 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.

- **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 country/economy as is feasible.
- **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 school-level exclusions.

**Pseudonymisation** – personal data where the personal identifier (e.g. respondent's name) is replaced by an artificial identifier (pseudonym). In PISA, the respondent's name is not included in the data collected on the assessment day, but is replaced with an ID number. Therefore the data is categorised as pseudonymised, as long as the additional information linking the respondent name and the PISA respondent ID is stored in a different file or location.

**School Associate (SA)** – a person at a school who acts as a liaison between the school and the National Centre to prepare for the assessment and who administers the assessment to students on the day of the assessment.

**School Co-ordinator (SC)** – a person at a school who acts as a liaison between the school and the National Centre to prepare for the assessment in the school.

**School-level exclusions** – contractors' approved 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, or
- of other reasons as agreed upon.

**School-level materials** – the key materials include:

- School Co-ordinator Manual and Test Administrator Manual (or School Associate Manual)
- Test administration scripts
- Key forms – Student Tracking Form, Session Attendance Form, and Session Report Form

**Source versions** assessment instruments provided in English (and, for some documents, in French) by the international contractors according to contractual specifications.

**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.

**Test administrator** – a person who is trained by the National Centre to administer the PISA test in schools. This person may be a Test Administrator or a School Associate (a School Co-ordinator who also has the role of a Test Administrator).

**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** – potential exclusion of students from 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, and
  - 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 J. PISA 2022 Contractors, Staff and Consultants

PISA is a collaborative effort, bringing together experts from the participating countries, steered jointly by their governments based on shared, policy-driven interests.

A PISA Governing Board, on which each country is represented, determines the policy priorities for PISA, in the context of OECD objectives, and oversees 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 2022, the overall management of contractors and implementation was carried out by the Educational Testing Service (ETS) in the United States as part of its responsibility as the Core A contractor. The OECD Secretariat worked closely with the International Project Director and Project Manager, to co-ordinate all aspects of implementation. In addition to overall management, Core A was responsible for the computer-delivery platform, instrument development, scaling and analysis, and all data products. As the lead of Core A, ETS worked in co-operation with Westat in the United States for survey operations, cApStAn for translation and verification of the assessment instruments, the International Association for Evaluation of Educational Achievement (IEA) in the Netherlands for the data management software,

The additional tasks related to the implementation of PISA 2022 were carried out by three additional contractors – Cores B1, B2, B3, C, D, and E.

The Research Triangle Institute (RTI) in the United States facilitated the development of the mathematics assessment framework as the Core B1 contractor. ETS also facilitated the development of the background questionnaire frameworks as the Core B2 contractor. ACT in the United States and Cito in the Netherlands performed the test development for the innovative domain as the Core B3 contractor. Core C focused on sampling and was implemented by Westat in the United States in co-operation with the Australian Council for Educational Research (ACER). Core D was managed by cApStAn Linguistic Quality Control in Belgium for linguistic quality control in co-operation with BranTra in Belgium. Core E focused on country preparation and implementation support and was managed by the Australian Council for Educational Research (ACER) in Australia.

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 representative who was involved in PISA 2022)

**Chair of the PISA Governing Board: Michele Bruniges**

### ***OECD Members and PISA Associates***

Australia: Meg Brighton, Alex Gordon\*, Ros Baxter\*, Rick Persse\*, Gabrielle Phillips\*

Austria: Mark Németh

Belgium: Isabelle Erauw, Geneviève Hindryckx

Brazil: Manuel Fernando Palacios Da Cunha E Melo, Carlos Eduardo Moreno Sampaio\*, Manuel Palácios\*, Danilo Dupas Ribeiro\*, Alexandre Ribeiro Pereira Lopes\*, Elmer Coelho Vicenzi\*, Marcus Vinícius Carvalho Rodrigues\*, Maria Inês Fini\*

Canada: Bruno Rainville, Manuel Cardoso\*, Kathryn O'Grady\*, Gilles Bérubé\*, Tomasz Gluszynski\*

Chile: Claudia Matus

Colombia: Elizabeth Blandon, Luisa Fernanda Trujillo Bernal \*, Andrés Elías Molano Flechas\*, Mónica Ospina Londoño\*, María Figueroa Cahnspeyer\*, Arango María Sofía\*

Costa Rica: Alvaro Artavia Medriano, Melvin Chaves Duarte, María Ulate Espinoza\*, Lilliam Mora\*, Melania Brenes Monge\*, Pablo José Mena Castillo\*, Edgar Mora Altamirano\*

Czech Republic: Tomas Zatloukal

Denmark: Hjalte Meilvang, Eydun Gaard, Charlotte Rotbøll Sjøgreen\*, Cecilie Kynemund\*, Frida Poulsen\*

Estonia: Maie Kitsing

Finland: Tommi Karjalainen, Najat Ouakrim-Soivio\*

France: Ronan Vourc'h, Sandra Andreu, Thierry Rocher\*

Germany: Jens Fischer-Kottenstede, Kathrin Stephen, Katharina Koufen\*, Elfriede Ohrnberger\*

Greece: Chryssa Sofianopoulou, Ioannis Tsirmpas\*

Hungary: Sándor Brassói

Iceland: Sigridur Lara Asbergisdóttir, Stefán Baldursson\*

Ireland: Rachel Perkins, Caroline McKeown\*

Israel: Gal Alon, Hagit Glickman\*

Italy: Roberto Ricci

Japan: Akiko Ono, Yu Kameoka\*

Korea: Kija Si, Hee Seung Yuh, Yun Jung Choi\*, Younghoon Ko\*, HeeKyoung Kim\*, Jeik Cho\*, Jimin Cho\*, Ji-young Park\*, Bae Dong-in\*

Latvia: Aļona Babiča

Lithuania: Rita Dukynaite

Mexico: Roberto Pulido, Antonio Ávila Díaz\*, Andrés Eduardo Sánchez Moguel\*, Bernardo H. Naranjo\*

Netherlands: Schel Margot, Marjan Zandbergen\*

New Zealand: Grant Pollard, Tom Dibley\*, Alex Brunt\*, Philip Stevens\*, Craig Jones\*

Norway: Marthe Akselsen

Poland: Piotr Mikiewicz

Portugal: Luís Pereira Dos Santos

Slovak Republic: Ivana Pichanicova, Romana Kanovská\*

Slovenia: Mojca Štraus, Ksenija Bregar Golobic

Spain: Carmen Tovar Sanchez

Sweden: Maria Axelsson, Ellen Almgren\*

Switzerland: Peter Lenz, Camil Würgler, Reto Furter\*, Vera Husfeldt\*

Thailand: Thiradet Jiarasuksakun, Supattra Pativisan, Nantawan Somsook\*, Sukit Limpijumnong\*

Türkiye: Umut Erkin Taş, Murat İlikhan\*, Sadri Şensoy\*, Kemal Bülbül\*

United Kingdom: Ali Pareas, Keith Dryburgh, Lorna Bertrand\*

United States: Peggy Carr

### ***Observers (Partner economies)***

Albania: Zamira Gjini

Argentina: Paula Viotti, Bárbara Briscioli\*, María Angela Cortelezzi\*, Elena Duro\*

Azerbaijan: Elnur Aliyev, Narmina Huseynova\*, Emin Amrullayev\*

Brunei Darussalam: Shamsiah Zuraini Kanchanawati Tajuddin, Hj Azman Bin Ahmad\*

Bulgaria: Neda Oscar Kristanova

Cambodia: Kreng Heng, Samith Put\*

Chinese Taipei: Yuan-Chuan Cheng, Chung-Hsi Lin\*, Tian-Ming Sheu\*

Croatia: Marina Markuš Sandric, Ines Elezović\*

Dominican Republic: Ancell Scheker Mendoza

El Salvador: Martin Ulises Aparicio Morataya, Óscar de Jesús Águila Chávez\*

Georgia: Sophia Gorgodze

Guatemala: Marco Antonio Sáiz Choxim, Luisa Fernanda Müller Durán\*

Hong Kong, China: Chi-fung Hui, Wai-sun Lau, Man-keung Lau\*, Hiu-fong Chiu\*, Ho Pun Choi\*

Indonesia: Anindito Aditomo, Totok Suprayitno\*



Jamaica: Terry-Ann Thomas-Gayle

Jordan: Abdalla Yousef Awad Al-Ababneh

Kazakhstan: Magzhan Amangazy, Miras Baimyrza\*, Yerlikzhan Sabyruly\*, Magzhan Amangazy\*, Yerlikzhan Sabyruly\*

Kosovo: Shqipe Bruqi, Agim Berdyna\*, Valmir Gashi\*

Lebanon: Hyam Ishak, Bassem Issa, George Nohra\*, Nada Oweijane\*

Macau, China: Chi Meng Kong, Kin Mou Wong, Pak Sang Lou\*

Malaysia: Ahmad Rafee Che Kassim, Pkharuddin Ghazali\*, Hajah Roziah Binti Abdullah\*, Habibah Abdul Rahim\*

Malta: Charles L. Mifsud

Republic of Moldova: Anatolie Topală

Mongolia: Oyunaa Purevdorj, Nyam-Ochir Tumor-Ochir\*, Tumorkhoo Uuganbayar\*

Montenegro: Miloš Trivic, Dragana Dmitrovic\*

Morocco: Youssef El Azhari, Mohammed Sassi\*

Republic of North Macedonia: Biljana Mihajloska, Natasha Jankovska\*, Natasha Janevska\*

Palestinian Authority: Mohammad Matar

Panama: Gina Garcés, Nadia De Leon\*

Paraguay: Sonia Mariángeles Domínguez Torres, Karen Edith Rojas de Riveros\*

People's Republic of China: Xiang Mingcan, Zhang Jin\*

Peru: Tania Magaly Pacheco Valenzuela, Gloria María Zambrano Rozas\*, Humberto Perez León Ibáñez\*

Philippines: Gina Gonong, Alma Ruby C. Torio\*, Jose Ernesto B. Gaviola\*, Diosdado San Antonio\*, Nepomuceno A. Malaluan\*

Qatar: Khalid Abdulla Q. Al-Harqan

Romania: Bogdan Cristescu, Daniela Elisabeta Bogdan\*

Saudi Arabia: Abdullah Alqataee, Husam Zaman\*, Faisal bin Abdullah Almishari Al Saud\*

Serbia: Branislav Randjelovic, Anamarija Viček\*

Singapore: Chern Wei Sng

Ukraine: Sergiy Rakov

United Arab Emirates: Hessa Al Wahabi, Rabaa Alsumaiti\*

Uruguay: Adriana Aristimuno, Andrés Peri\*

Uzbekistan: Abduvali Abdumalikovich Ismailov, Radjiyev Ayubkhon Bakhtiyorkhonovich\*

Viet Nam: Huynh Van Chuong, Le My Phong\*, Sai Cong HONG\*

## PISA 2022 National Project Managers

(\*Former PISA 2022 NPM)

### **OECD Members and PISA Associates**

Australia: Lisa De Bortoli, Sue Thomson\*

Austria: Birgit Lang, Bettina Toferer

Belgium: Inge De Meyer, Anne Matoul

Brazil: Clara Machado Da Silva Alarcão, Aline Mara Fernandes Muler, Katia Pedroza\*, Wallace Nascimento Pinto Junior\*

Canada: Vanja Elez, Kathryn O'Grady\*, Tanya Scerbina\*

Chile: Ema Lagos Campos

Colombia: Julie Paola Caro Osorio, Natalia González Gómez\*

Costa Rica: Rudy Masís Siles, Giselle Cruz Maduro\*

Czech Republic: Simona Boudova, Radek Blazek\*

Denmark: Vibeke Tornhøj Christensen, Ása Hansen, Magnus Bjørn Sørensen\*

Estonia: Gunda Tire

Finland: Arto Ahonen, Mari-Pauliina Vainikainen

France: Franck Salles, Irène Verlet\*

Germany: Jennifer Diedrich-Rust, Doris Lewalter, Mirjam Weis, Kristina Reiss\*

Greece: Chryssa Sofianopoulou

Hungary: Csaba Rózsa, Judit Szipocs-Krolopp, László Ostorics\*

Iceland: Guðmundur Þorgrímsson

Ireland: Brenda Donohue

Israel: Georgette Hilu, Inbal Ron-Kaplan

Italy: Carlo Di Chiacchio, Laura Palmerio

Japan: Naoko Otsuka, Kentaro Sugiura\*, Yu Kameoka\*,

Korea: Seongkyeong Kim, Shinyoung Lee\*, Inseon Choi\*, Seongmin Cho\*

Latvia: Rita Kiseļova

Lithuania: Rasa Jakubauske, Natalija Valaviciene\*, Mindaugas Stundža\*

Mexico: Proceso Silva Flores, Rafael Vidal\*, Mariana Zuniga Garcia\*, María Antonieta Díaz Gutierrez\*

Netherlands: Joyce Gubbels, Martina Meelissen

New Zealand: Steven May, Emma Medina, Adam Jang-Jones\*

Norway: Fredrik Jensen

Poland: Krzysztof Bulkowski, Joanna Kazmierczak

Portugal: Anabela Serrão

Slovak Republic: Júlia Miklovičová

Slovenia: Klaudija Šterman Ivancic

Spain: Lis Cercadillo

Sweden: Maria Axelsson

Switzerland: Andrea Erzinger

Thailand: Ekarin Achakunwisut

Türkiye: Umut Erkin Taş

United Kingdom: Grace Grima, David Thomas, Juliet Sizmur\*

United States: Samantha Burg, Patrick Gonzales\*

### ***Observers (Partner economies)***

Albania: Aurora Balliu, Rezana Vrapic\*

Argentina: Maria Clara Radunsky, Paula Viotti\*, Raul Volker\*, Cecilia Beloqui\*

Azerbaijan: Ulkar Zaidzada, Zinyat Amirova\*, Leyla Abbasli\*

Brunei Darussalam: Wan Abdul Rahman Wan Ibrahim, Hazri Haji Kifle\*

Bulgaria: Natalia Vassileva

Cambodia: Chinna Ung

Chinese Taipei: Chin-Chung Tsai

Croatia: Ana Markočić Dekanić

Dominican Republic: Santa Cabrera, Claudia Curiel\*, Massiel Cohen Camacho\*

El Salvador: José Carlos Márquez Hernández

Georgia: Tamari Shoshitashvili, Lasha Kokilashvili\*

Guatemala: Marco Antonio Sáiz Choxim, Luisa Fernanda Müller Durán\*

Hong Kong, China: Kit-Tai Hau

Indonesia: Asrijanty Asrijanty, Moch Abduh\*

Jamaica: Marjoriana Clarke

Jordan: Emad Ghassab Ababneh

Kazakhstan: Rizagul Syzdykbayeva, Nadezhda Cherkashina\*

Kosovo: Fatmir Elezi

Lebanon: George Nohra, Nada Oweijane\*

Macau, China: Kwok-Cheung Cheung

Malaysia: Wan Faizatul Shima Ismayatim, Wan Raisuha Wan Ali, Hajah Roziah Binti Abdullah\*, Azhar Ahmad\*, Ahmad Rafee Che Kassim\*

Malta: Jude Zammit, Louis Scerri\*

Republic of Moldova: Anatolie Topală

Mongolia: Tungalagtuul Khaltar

Montenegro: Divna Paljevic

Morocco: Anass El Asraoui, Ahmed Chaibi

Republic of North Macedonia: Beti Lameva

Palestinian Authority: Mohammad Matar

Panama: Arafat A. Sleiman G., Ariel Melo\*

Paraguay: Judith Franco Ortega, Verónica Heilborn Díaz\*

People's Republic of China: Tao Xin

Peru: Tania Magaly Pacheco Valenzuela, Gloria María Zambrano Rozas\*, Humberto Perez León Ibáñez\*

Philippines: Nelia V. Benito

Qatar: Shaikha Al-Ishaq

Romania: Gabriela Nausica Noveanu, Petre Feodorian Botnariuc\*

Saudi Arabia: Abdullah Aljouiee, Fahad Ibrahim Almoqhim\*

Serbia: Gordana Čaprić

Singapore: Elaine Chua

Ukraine: Tetiana Vakulenko

United Arab Emirates: Shaikha Alzaabi, Ahmed Hosseini, Moza Rashid Ghufli\*

Uruguay: Laura Noboa, María H. Sánchez\*

Uzbekistan: Abduvali Abdumalikovich Ismailov

Viet Nam: Quoc Khanh Pham, Thi My Ha Le\*

## OECD Secretariat

Andreas Schleicher (Strategic development)

Francesco Avvisati (Analysis and reporting, and Research, Development and Innovation)

Charlotte Baer (Communications)

Anna Becker (Research, Development and Innovation)

Yuri Belfali (Strategic development)

Guillaume Bousquet (Analysis and reporting)

Janine Buchholz (Research, Development and Innovation)

Eda Cabbar (Production support)

Tiago Caliço (Research, Development and Innovation)

Rodrigo Castaneda Valle (Analysis and reporting)

Marta Cignetti (Research, Development and Innovation)

Catalina Covacevich (Project management)

Duncan Crawford (Communications)  
Alfonso Echazarra (Analysis and reporting)  
Natalie Foster (Research, Development and Innovation)  
Tiago Fragoso (Project management)  
Marc Fuster Rabella (Research, Development and Innovation)  
Kevin Gillespie (Communications, and Project management)  
Juliana Andrea González Rodríguez (Project management)  
Ava Guez (Research, Development and Innovation)  
Tue Halgreen (Project management)  
Kartika Herscheid (Analysis and reporting)  
Irène Hu (Analysis and reporting)  
Miyako Ikeda (Analysis and reporting)  
Gwénaél Jacotin (Analysis and reporting)  
Kristina Jones (Project management)  
Theo Kaiser (Research, Development and Innovation)  
Natalie Laechelt (Project management)  
Gracelyn Lee (Analysis and reporting)  
Emma Linsenmayer (Research, Development and Innovation)  
Adrien Lorenceau (Analysis and reporting)  
Camille Marec (Analysis and reporting)  
Thomas Marwood (Project management)  
Caroline McKeown (Project management)  
Chiara Monticone (Analysis and reporting)  
Tarek Mostafa (Analysis and Reporting)  
Josephine Murasiranwa (Research, Development and Innovation)  
Lesley O'Sullivan (Project management)  
Valeria Pelosi (Analysis and reporting)  
Mario Piacentini (Research, Development and Innovation)  
Sasha Ramirez-Hughes (Communications)  
Giannina Rech (Analysis and reporting)  
Daniel Salinas (Analysis and reporting)  
Ricardo Sanchez Torres (Project management)  
Della Shin (Communications)  
Javier Suarez-Alvarez (Analysis and reporting)  
Lucia Tramonte (Analysis and reporting)

Chi Sum Tse (Project management)  
 Hannah Ulferts (Analysis and reporting)  
 Hanna Varkki (Project management)  
 Sophie Vayssettes (Project management)  
 Nathan Viltard (Analysis and reporting)  
 Michael Ward (Project Management)  
 Megan Welsh (Research, Development and Innovation)  
 Choyi Whang (Analysis and reporting)

### Mathematics Expert Group (MEG)

Joan Ferrini-Mundy (University of Maine, United States)  
 Zbigniew Marciniak (University of Warsaw, Poland)  
 William Schmidt (Michigan State University, United States)  
 Takuya Baba (Hiroshima University, Japan)  
 Jenni Ingram (University of Oxford, United Kingdom)  
 Julián Mariño (University of the Andes, Colombia)

### Extended Mathematics Expert Groups (eMEG)

Michael Besser (Leuphana University of Lüneburg, Germany)  
 Jean-Luc Dorier (University of Geneva, Switzerland)  
 Iddo Gal (University of Haifa, Israel)  
 Markku Hannula (University of Helsinki, Finland)  
 Hannes Jukk (University of Tartu, Estonia)  
 Christine Stephenson (University of Tennessee, United States)  
 Tin Lam Toh (Nanyang Technological University, Singapore)  
 Ödön Vancsó (Eötvös Loránd University, Hungary)  
 David Weintrop (College of Information Studies, University of Maryland, United States)  
 Richard Wolfe (Ontario Institute for Studies in Education, University of Toronto, Canada)

### Financial Literacy Expert Group (FLEG)

Carmela Aprea (University of Mannheim, Germany)  
 José Alexandre Cavalcanti Vasco (Securities and Exchange Commission, Brazil)  
 Paul Gerrans (University of Western Australia, Australia)  
 David Kneebone (Investor Education Centre, Hong Kong (China))

Sue Lewis (Financial Services Consumer Panel, United Kingdom)

Annamaria Lusardi (George Washington University School of Business and Global Financial Literacy Excellence Center, United States)

Olaf Simonse (Ministry of Finance, Netherlands)

Anna Zelentsova (Ministry of Finance of the Russian Federation, Russia)

### **Creative Thinking Expert Group (CTEG)**

Ido Roll (Technion - Israel Institute of Technology, Israel)

Baptiste Barbot (Université Catholique de Louvain, Belgium)

Lene Tanggaard (Aalborg University, Denmark)

Nathan Zoanetti (Australian Council for Educational Research, Australia)

James Kaufman (University of Connecticut, United States)

Marlene Scardamalia (University of Toronto, Canada)

Valerie Shute (Florida State University, United States)

### **Questionnaire Expert Group (QEG)**

Nina Jude (Heidelberg University, Germany)

Hunter Gehlbach (Johns Hopkins University, United States)

Kit-Tai Hau (The Chinese University of Hong Kong, Hong Kong (China))

Therese Hopfenbeck (University of Melbourne, Australia)

David Kaplan (University of Wisconsin-Madison, United States)

Jihyun Lee (University of New South Wales, Australia)

Richard Primi (Universidade São Francisco, Brazil)

Wilima Wadhwa (ASER Centre, India)

### **Questionnaire Senior Framework Advisors**

Jennifer Adams (Ottawa-Carleton School District, Canada)

Eckhard Klieme (German Institute for International Educational Research, Germany)

Reinhard Pekrun (University of Essex, United Kingdom)

Jennifer Schmidt (Michigan State University, United States)

Arthur Stone (University of Southern California, United States)

Roger Tourangeau (Westat, United States)

Fons J.R. van de Vijver (Tilburg University/North-West University/University of Queensland)

## ICT Expert Group

Michael Trucano (World Bank, United States)  
 Jepe Bundsgaard (University of Aarhus, Denmark)  
 Cindy Ong (Ministry of Education, Singapore)  
 Patricia Wastiau (European Schoolnet, Belgium)  
 Pat Yongpradit (Code.org, United States)

## Technical Advisory Group

Keith Rust (Westat, United States)  
 Kentaro Yamamoto (ETS, United States)  
 John de Jong\* (VU University Amsterdam, Netherlands)  
 Christian Monseur (University of Liège, Belgium)  
 Leslie Rutkowski (chair) (University of Oslo, Norway and Indiana University, United States)  
 Eugenio Gonzalez, Ann Kennedy\*, Claudia Tamassia\* (ETS, United States)  
 Oliver Lüdtke (IPN - Leibniz Institute for Science and Mathematics Education, Germany)  
 Kathleen Scalise (University of Oregon, United States)  
 Sabine Meinck (IEA, Hamburg, Germany)  
 Kit-Tai Hau (Chinese University of Hong Kong, Hong Kong, China)  
 Maria Bolsinova (Tilburg University, the Netherlands)  
 Matthias von Davier\* (NBME, United States)

## PISA 2022 Lead Contractors

\* Indicates formerly in the position.

### ***Educational Testing Service (United States) – Core A and Core B2 lead contractors***

Irwin Kirsch (International Project Director)  
 Eugenio Gonzalez, Ann Kennedy\*, Claudia Tamassia\* (International Project Manager)  
     Larry Hanover (Editorial Support)  
     Luisa Langan\* (Project Management, Questionnaires)  
     Judy Mendez (Project Support and Contracts)  
     J. Franco (Project Support)  
     Daniel Nicastro (Project Support)  
     Yelena Shuster\* (Project Support)  
 Kentaro Yamamoto\* (Director, Psychometrics and Analysis)



Fred Robin (Manager, Psychometrics and Analysis)

Usama Ali (Psychometrics and Analysis)

Selene Sunmin Lee (Psychometrics and Analysis)

Emily Lubaway (Psychometrics and Analysis)

Peter van Rijn (Psychometrics and Analysis)

Hyo Jeong Shin (Psychometrics and Analysis)

David Garber (Lead Test Developer and Test Development Coordinator, Mathematical Literacy)

Elisa Giaccaglia (Test Developer and Reviewer, Mathematical Literacy)

Jeff Haberstroh (Test Developer and Reviewer, Mathematical Literacy)

Alessia Marigo (Test Developer and Reviewer, Mathematical Literacy)

Brian Sucevic (Test Developer and Reviewer, Mathematical Literacy)

James Meadows (Reviewer, Mathematical Literacy)

Enruo Guo (Interface Design, Mathematical Literacy)

Janet Stumper (Graphic Design, Mathematical Literacy)

Michael Wagner (Director, Platform Development)

Jason Bonthron (Platform Development and Authoring)

Paul Brost (Platform Development)

Ramin Hemat (Platform Development and Authoring)

Keith Kiser (Platform Development and Coding System)

Debbie Pisacreta (Interface Design and Graphics)

Janet Stumper (Graphics)

Chia Chen Tsai (Platform Development)

Edward Kulick\* (Area Director, Data Analysis and Research Technologies)

Mathew Kandathil Jr. (Manager, Data Analysis and Research Technologies)

Carla Tarsitano (Project Management)

John Barone\* (Data Products)

Kevin Bentley (Data Products)

Hezekiah Bunde (Data Management)

Karen Castellano (Data Management)

Matthew Duchnowski\* (Data Management)

Ying Feng (Data Management)

Zhumei Guo (Data Analysis)

Paul Hilliard (Data Analysis)

Lokesh Kapur (Data Analysis)

Debra Kline\* (Project Management)

Phillip Leung (Data Products)  
 Alfred Rogers\* (Data Management)  
 Tao Wang (Data Products)  
 Lingjun Wong (Data Analysis)  
 Ping Zhai\* (Data Analysis)  
 Shuwen Zhang\* (Data Analysis)  
 Wei Zhao (Data Analysis)

Jonas Bertling (Director, Questionnaire Framework and Development)  
 Jan Alegre (Questionnaire Framework and Development)  
 Katie Faherty (Questionnaire Management and Development)  
 Janel Gill (Questionnaire Scaling and Analysis)  
 Farah Qureshi (Team Assistance)  
 Nate Rojas (Team Assistance)

***Research Triangle Institute (RTI) and Pearson – Core B1 contractors***

Kimberly O'Malley (Project Director)  
 Jason Hill, Dave Leach (Project Manager)  
 Aarnout Brombacher (Content Lead)  
 Ben Dalton (Administrative Support Staff)  
 Tracy Kline (Administrative Support Staff)  
 Wendi Ralaingita (Administrative Support Staff)  
 Yasmin Sitabkhan (Administrative Support Staff)

***ACT (United States) and CITO (the Netherlands) – Core B3 contractors***

Andrew Taylor, Yigal Rosen\*, Gunter Maris\*, Alina von Davier\* (Programme Director)  
 Matthew Lumb, Ken Kobell\* (Programme Manager)  
 Kristin Lansing-Stoeffler, Yigal Rosen\* (Assessment Design Lead)  
 Kurt Peterschmidt (Technical Design Lead)  
 Matt Lumb (Scoring Design Lead)  
 Iris Garcia (Scoring Design Support)  
 Nicole Johnson (Scoring Design Support)  
 Chi-Yu Huang, Gunter Maris\* (Data Analysis Lead)  
 Shalini Kapoor (Data Analysis Support)  
 NooRee Huh (Data Analysis Support)  
 Jeffrey Steedle (Data Analysis Support)

Ben Deonovic (Data Analysis Support)

Chakadee Boonkasame (Data Analysis Support)

Jeremy Burrus (Content Lead, Background Questionnaire)

Cristina Anguiano-Carrasco (Support for Background Questionnaire)

Kate Walton (Support for Background Questionnaire)

***WESTAT (United States) – Core C lead contractor***

Keith Rust (Director of the PISA Consortium for Sampling and Weighting)

Sheila Krawchuk (Sampling)

David Ferraro (Sampling and Weighting)

Josephine Auguste (Weighting)

Jill DeMatteis (Sampling and Weighting)

Shaohua Dong (Sampling)

Susan Fuss (Sampling and Weighting)

Moriah Goodnow (Sampling and Weighting)

Amita Gopinath (Weighting)

Daniel Guzman (Sampling)

Jing Kang (Sampling and Weighting)

Sihle Khanyile (Weighting)

Véronique Lieber (Sampling and Weighting)

John Lopdell (Sampling and Weighting)

Shawn Lu (Weighting)

Irene Manrique Molina (Sampling and Weighting)

Leanna Moron (Sampling and Weighting)

Jacqueline Severynse (Sampling and Weighting)

Yumiko Siegfried (Sampling and Weighting)

Joel Wakesberg (Sampling and Weighting)

Sipeng Wang (Sampling and Weighting)

Natalia Weil (Sampling and Weighting)

Erin Wiley (Sampling and Weighting)

Sergey Yagodin (Weighting)

***cApStAn Linguistic Quality Control (Belgium) – Core D lead contractor***

Steve Dept (Project Director, Translatability Assessment)

Andrea Ferrari (Linguistic Quality Assurance and Quality Control Designs)

Emel Ince (Verification Management, Coding Guides)

Elica Krajičeva (Lead Project Manager)

Shinoh Lee (Verification Management, Questionnaires)

Irene Liberati (Verification Management, Cognitive Units, Coding Guides)

Roberta Lizzi (Verification Management, Cognitive Units)

Adrien Mathot (Translation Technologist, Linguistic Quality Assurance Tools and Procedures)

Manuel Souto Pico (Lead Translation Technologist, Linguistic Quality Assurance Tools and Procedures)

Josiane Tyburn (Verification Management, Questionnaires, School Materials)

### ***Australian Council for Educational Research (Australia) – Core E lead contractor***

Jeaniene Spink, Maurice Walker (Project Directors)

Jennie Chainey

Jacqueline Cheng

Sandra Lambey

Naoko Tabata

Ursula Schwantner

## **PISA 2022 Contributors, working with Lead Contractors**

### ***Australian Council for Educational Research (Australia) – Core C contributor***

Martin Murphy (Project Director)

Emma Cadman (School Sampling)

Emma Camus (School Sampling)

Martin Chai (Student Sampling)

Alex Daragonov (Student Sampling)

Jorge Fallas (Lead School Sampling)

Kathy He (Student Sampling)

Jennifer Hong (School and Student Sampling)

Yan Jiang (Student Sampling)

Renee Kwong (School and Student Sampling)

Dulce Lay (School Sampling)

Nina Martinus (School Sampling)

Louise Ockwell (Student Sampling)

Claire Ozolins (School Sampling)

Anna Plotka (Student Sampling)

Alla Routitsky (Student Sampling)

Paul Tabet (School Sampling)

***BranTra (Belgium) – Core D contributor***

Eva Jacob (Translation Management, French Source Development)

Danina Lupsa (Translation Technologist, Linguistic Quality Assurance Tools and Procedures)

Ben Meessen (Translation Management, Development of Common Reference Versions for Spanish, Chinese, Arabic)

***HallStat SPRL (Belgium) – Core A contributor as the translation referee***

Beatrice Halleux (Consultant, Translation/Verification Referee, French Source Development)

***WESTAT (United States) – Core A contributor on survey operations***

Merl Robinson (Director of Core A Contractor for Survey Operations)

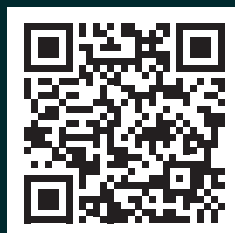
Michael Lemay (Manager of Core A Contractor for Survey Operations)

Sarah Sparks (National Centre Support, Quality Control)

Beverley McGaughan (National Centre Support, Quality Control)

# PISA 2022 Technical Report

The Programme for International Student Assessment (PISA) is one of the largest and most comprehensive comparative education studies in the world. A wide variety of countries and economies worldwide collect information on student performance, school environment, and other relevant variables using standardized, uniform procedures that assure the results are comparable and meaningful. This Technical Report has been prepared by those who implemented PISA during its 2022 cycle to provide transparency to these procedures and to the statistical and mathematical methods that underpin the comparability and validity of PISA 2022 results.



PRINT ISBN 978-92-64-92890-9  
PDF ISBN 978-92-64-82476-8

