

*IEA International Computer and
Information Literacy Study 2023*

TECHNICAL REPORT

(Revised)

Julian Fraillon
Mojca Rožman
Sebastian Meyer
Lauren Musu
Yuan-Ling Liaw
Andrés Christiansen
Sabine Tieck

Editors

Co-funded by the



International Computer and
Information Literacy Study 2023

Technical Report

Editors

Julian Fraillon • Mojca Rožman • Sebastian Meyer
Lauren Musu • Yuan-Ling Liaw • Andrés Christiansen
Sabine Tieck

International Computer and
Information Literacy Study 2023

Technical Report



IEA
Keizersgracht 311
1016 EE Amsterdam
The Netherlands
Telephone: +31 20 625 3625
Fax: + 31 20 420 7136
Email: secretariat@iea.nl
Website: www.iea.nl



The International Association for the Evaluation of Educational Achievement (IEA) is an independent, international cooperative of national research institutions and governmental research agencies. It conducts large-scale comparative studies of educational achievement and other aspects of education, with the aim of gaining in-depth understanding of the effects of policies and practices within and across systems of education.

Cover design by Studio Lakmoes, Arnhem, The Netherlands.

© Stichting IEA Secretariaat Nederland 2025.
This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

This work is subject to copyright. All commercial rights are reserved by the author(s), whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Regarding these commercial rights a non-exclusive license has been granted to the publisher.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use. The publisher, the authors and the editors are safe to assume that the advice and information in this book are to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Contents

1	Introduction	1
	Julian Fraillon	
	1.1 Background	1
	1.2 Study overview	2
	1.3 Outline of this technical report	7
	References	7
2	ICILS 2023 test development	9
	Daniel Duckworth and Julian Fraillon	
	2.1 Introduction	9
	2.2 Terminology	9
	2.3 Test scope and format	10
	2.4 Test-development process	15
	2.5 Field trial test design and content	20
	2.6 Main survey test design and content	25
	2.7 Released test modules	30
	References	30
3	ICILS survey operations software	31
	Sebastian Meyer and Daniel Duckworth	
	3.1 Introduction	31
	3.2 Survey operations software applications and systems	31
	3.3 Procedures and software user manuals	32
	3.4 ICILS 2023 system architecture	33
	3.5 Instrument preparation	33
	3.6 Student instrument delivery using AM Examiner	36
	3.7 Principal, ICT coordinator, and teacher questionnaires delivery using OSS web component	40
	3.8 Student response scoring and coding	42
	3.9 Managing the online systems and data collection	43
	3.10 Challenges with computer-based delivery in ICILS	44
	References	46
4	ICILS questionnaire development	47
	Mojca Rožman and Julian Fraillon	
	4.1 Introduction	47
	4.2 The conceptual framework and questionnaire content	47
	4.3 Development process	49
	References	56
5	Preparation and verification of the ICILS 2023 study instruments	57

Marta Moreno Hidalgo and Tim Daniel

5.1	Description of the materials, languages, and systems	58
5.2	Adaptation, translation, and layout reviews	59
5.3	International verification processes	65
6	Sampling design and implementation	71
	Sabine Tieck	
6.1	Introduction	71
6.2	Target population definitions	71
6.3	Coverage and exclusions	73
6.4	School sampling design	80
6.5	Within-school sampling design	85
6.6	Sample size requirements	85
	References	88
7	Sampling weights and participation rates	89
	Sabine Tieck	
7.1	Introduction	89
7.2	Types of sampling weights	89
7.3	Calculating weights for the student population	90
7.4	Calculating weights for the teacher population	92
7.5	Calculating weights for the school-level survey	94
7.6	Calculating participation rates	94
7.7	ICILS 2023 standards for sampling participation	102
7.8	Reporting data	105
	References	112
8	ICILS field operations	113
	Sebastian Meyer	
8.1	Introduction	113
8.2	Field operations personnel	113
8.3	Field operations resources	114
8.4	Field operations processes	115
8.5	Field trial procedures	124
9	Quality assurance in the ICILS data collection	125
	Kateřina Hartmanov, Lauren Musu, and Marta Moreno Hidalgo	
9.1	International quality control program	125
9.2	Findings from the ICILS 2023 administration observations	127
9.3	Findings from the national materials review	143
9.4	Findings from the survey activities questionnaire	144
10	Data management and creation of the ICILS database	157
	Tim Daniel	
10.1	Introduction	157
10.2	Data sources	157
10.3	Confirming the integrity of the national databases	159

10.4 The ICILS 2023 international database	166
10.5 Summary	167
References	167
11 Scaling procedures for ICILS 2023 test items	169
Yuan-Ling Liaw, Mojca Rožman, Julian Fraillon, and Eugenio Gonzalez	
11.1 Introduction	169
11.2 Overview of scaling methodology	170
11.3 Implementation of scaling procedures: test and item adjudication	173
11.4 Implementation of scaling procedures: item calibration	192
11.5 Implementation of scaling procedures: population modeling to estimate students' scores	197
11.6 Proficiency levels for CIL and CT	203
References	206
12 Scaling procedures for ICILS 2023 questionnaire items	207
Minge Chen, Andrés Christiansen, Mojca Rožman, and Rolf Strietholt	
12.1 Simple indices	207
12.2 Students' socioeconomic background indices	213
12.3 Scaling procedures	215
12.4 Scaling results	218
References	264
13 Reporting the ICILS 2023 results	267
Andrés Strello, Mojca Rožman, and Julian Fraillon	
13.1 Estimation of sampling variance	267
13.2 Estimation of imputation variance for CIL and CT scores	270
13.3 Reporting relationships between CIL and CT scores	273
13.4 Reporting of differences	274
13.5 ICILS 2023 reporting policy	279
13.6 Notes about software	280
References	280
Appendices	282
A Organizations and individuals involved in ICILS 2023	283
B National characteristics	289
B.1 Austria	289
B.2 Azerbaijan	290
B.3 Belgium (Flemish)	291
B.4 Bosnia and Herzegovina	292
B.5 Chile	293
B.6 Chinese Taipei	294
B.7 Croatia	296
B.8 Cyprus	297
B.9 Czech Republic	298
B.10 Denmark	300
B.11 Finland	301

B.12	France	303
B.13	Germany	304
B.14	Greece	305
B.15	Hungary	307
B.16	Italy	308
B.17	Kazakhstan	309
B.18	Korea, Republic of	311
B.19	Kosovo	312
B.20	Latvia	313
B.21	Luxembourg	314
B.22	Malta	315
B.23	The Netherlands	316
B.24	Norway	317
B.25	Oman	318
B.26	Portugal	320
B.27	Romania	321
B.28	Serbia	323
B.29	Slovak Republic	324
B.30	Slovenia	325
B.31	Spain	326
B.32	Sweden	328
B.33	United States	329
B.34	Uruguay	330
B.35	Benchmark country: North-Rhine Westphalia (Germany)	331
C	CIL release module scoring rules	333
C.1	Breathing	333
C.2	School trip	342
D	CT release module scoring rules	349
D.1	Automated bus	349
D.2	Farm drone	353
	References	359
E	Background variables used for conditioning	361
F	National item adjudication	367
G	Recoding and scoring rules for variables included in scaling	369
G.1	Breathing	369
G.2	School trip	369
G.3	Automated bus	370
G.4	Farm drone	371
H	Details on questionnaire scale results	375
	Errata	391

List of Figures

4.1.	Contexts for ICILS 2023 CIL/CT learning outcomes	48
5.1.	ICILS 2023 national instrument preparation workflow	58
5.2.	ICILS 2023 adaptation types	61
5.3.	Track changes after translation verifier edits	67
6.1.	Target Population - Internationals desired to National defined	74
6.2.	Visualization of PPS systematic sampling.	84
7.1.	Participation categories in ICILS	105
8.1.	Procedures for working with schools to prepare for test administration	118
10.1.	Overview of data processing at the ISC	162
11.1.	Item characteristic curve	171
11.2.	Mapping of CIL student abilities, item difficulties, and country means	174
11.3.	Mapping of CT student abilities, item difficulties, and country means	175
11.4.	Item statistics for item R02Z	185
11.5.	Item statistics for item R02Z (cont'd)	186
11.6.	Item difficulties for common CIL (left) and CT (right) items in 2018 and 2023	193

List of Tables

1.1.	ICILS 2023 participating education systems	5
1.2.	Grade level equivalence and student age across countries	6
2.1.	Summary of ICILS 2023 main survey CIL test modules	12
2.2.	Summary of ICILS 2023 main survey CT test modules	13
2.3.	Test development processes and timeline	15
2.4.	Field trial CIL test module composition: Number of scoring variables by task type	21
2.5.	Field trial CT test module composition: Number of scoring variables by task type	22
2.6.	Field trial test form design and contents	23
2.7.	Field trial CIL scoring variable mapping to the CIL framework	24
2.8.	Field trial CT scoring variable mapping to the CT framework	25
2.9.	Main survey CIL test module composition: Number of scoring variables by task type	26
2.10.	Main survey CT test module composition: Number of scoring variables by task type	26
2.11.	Main survey test form design and contents	27
2.12.	Main survey CIL scoring variable mapping to the CIL framework	29
2.13.	Main survey CT scoring variable mapping to the CT framework	29
3.1.	ICILS survey operations software applications and systems	33
3.2.	Unweighted participation status percentages across participants based on full database and original coding by test administrators (reference: all sampled students)	45
4.1.	Mapping of variables to contextual framework (examples)	48
4.2.	Detailed process of questionnaire development	50
5.1.	Instruments that required translation/adaptation	58
5.2.	ICILS 2023 administered languages	60
5.3.	Trend countries (administration in 2018)	63
6.1.	Percentages of schools excluded from the ICILS target populations	76
6.2.	Percentages of students covered and excluded from the ICILS student target population	78
6.3.	Stratification schemes of participating countries	82

6.3.	Stratification schemes of participating countries (cont'd)	83
6.4.	School, student, and teacher sample sizes - expected versus realized	87
6.4.	School, student, and teacher sample sizes - expected versus realized (cont'd)	88
7.1.	Unweighted school, class, and student participation rate (PR) - student survey	97
7.2.	Weighted school, class, and student participation rate (PR) - student survey	98
7.3.	Unweighted school and teacher participation rate (PR) - teacher survey	101
7.4.	Weighted school and teacher participation rate (PR) - teacher survey	103
7.5.	Used Annotations for the student and teacher survey	106
7.6.	Reporting category of each country for the student survey	107
7.6.	Reporting category of each country for the student survey (cont'd)	108
7.7.	Reporting category of each country for the teacher survey	109
7.7.	Reporting category of each country for the teacher survey (cont'd)	110
7.7.	Reporting category of each country for the teacher survey (cont'd)	111
8.1.	Hierarchical identification codes	116
8.2.	Timing of the ICILS assessment	121
9.1.	Quality of ICILS 2023 administration preparation - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	128
9.2.	CIL Test Administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	129
9.3.	Break after CIL test administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	130
9.4.	Student questionnaire administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	131
9.5.	Break after student questionnaire administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	131
9.6.	CT module administration - Across 355 observed ICILS administration sessions where CT was administered (percent of IQCM responses, rounded to a whole number)	132
9.7.	General IQCM impressions and observations - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	133
9.8.	Student cooperation - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	134
9.9.	Special accommodations - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	135
9.10.	Technical issues and support - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)	136
9.11.	Administering the main survey - Accommodations (number of NRC responses, max 35)	137
9.12.	Interview with the school coordinator - Preparations (percent of school coordinator responses)	139
9.13.	Interview with the school coordinator - Listing forms (percent of responses)	140
9.14.	Interview with the school coordinator - test administrator training (percent of school coordinator responses)	140
9.15.	Interview with the school coordinator - Final observations (percent of responses)	141
9.16.	Interview with the test administrator (percent of test administrator responses)	142
9.17.	Sampling (number of NRC responses, max 35)	145
9.18.	Sampling - Procedure difficulty (number of NRC responses, max 35)	146
9.19.	Contacting schools and recruiting school coordinators (number of NRC responses, max 35)	146
9.20.	Training school coordinators (number of NRC responses, max 35)	147
9.21.	Adapting and translating the ICILS assessment materials (number of NRC responses, max 35)	147

9.22. Translator/reviewer competencies (number of NRC responses, max 35)	148
9.23. Making and documenting national adaptations using the national adaptations form (number of NRC responses, max 35)	148
9.24. Adapting and translating the ICILS assessment materials - Procedure difficulty (number of NRC responses, max 35)	149
9.25. Adapting and translating the ICILS assessment materials - Issues encountered (number of NRC responses, max 35)	149
9.26. Preparing the ICILS materials for administration - Issues encountered (number of NRC responses, max 35)	150
9.27. Administering the main survey - Test administrators (number of NRC responses, max 35)	150
9.28. Administering the main survey - Participation issues (number of NRC responses, max 35)	151
9.29. Administering the main survey - Accommodations (number of NRC responses, max 35)	152
9.30. National quality monitoring program (number of NRC responses, max 35)	153
9.31. Scoring CIL and CT test responses (number of NRC responses, max 35)	154
9.32. Entering and coding occupation data (number of NRC responses, max 35)	155
9.33. Collecting and submitting data (number of NRC responses, max 35)	155
9.34. The NRCs' assessment of the quality of the ICILS manuals (number of NRC responses, max 35)	156
11.1. Percentages of omitted responses and items not reached for CIL items	179
11.2. Percentages of omitted responses and items not reached for CIL items, by item type and module	182
11.3. Percentages of omitted responses and items not reached for CIL items by module	182
11.4. Percentages of omitted responses and items not reached for CT items	183
11.5. Percentages of omitted responses and items not reached for CT items, by item type and module	184
11.6. Scorer agreement percentages for constructed-response CIL test items	190
11.7. Scorer agreement percentages for constructed-response CT test items	192
11.8. Item parameter estimates of the CIL test items from concurrent calibration	194
11.9. Item parameter estimates of the CT test items from concurrent calibration	196
11.10. Conditioning models for CIL and CT achievement scales	198
11.11. ICILS 2018 published scores and re-scaled scores	201
11.12. Transformation constants for the ICILS 2023	201
11.13. Trend linking error variance between ICILS 2018 and ICILS 2023 calibrations in the CIL achievement scale	202
11.14. Trend linking error variance between ICILS 2018 and ICILS 2023 calibrations in the CT achievement scale	203
12.1. Percentage of imputed data, reliability, explained variance, and component loadings of NISB by country and IISB	214
12.2. Item wording for the scale Learning about internet related tasks at school (S_LRNINTS)	219
12.3. Item parameters, item fits, and WRMSD for the scale Learning about internet related tasks at school (S_LRNINTS)	219
12.4. Item wording for the scale Learning about internet related tasks outside of school (S_LRNINTO)	220
12.5. Item parameters, item fits, and WRMSD for the scale Learning about internet related tasks outside of school (S_LRNINTO)	220
12.6. Item wording for the scale Learning about safe and responsible ICT use at school (S_LRNSAFE)	221
12.7. Item parameters, item fits, and WRMSD for the scale Learning about safe and responsible ICT use at school (S_LRNSAFE)	221
12.8. Item wording for the scale Academic-media multitasking (S_ACMULT)	221

12.9. Item parameters, item fits, and WRMSD for the scale Academic-media multitasking (S_ACMULT)	222
12.10.Item wording for the scale General classroom applications (S_GENCLASS)	222
12.11.Item parameters, item fits, and WRMSD for the scale General classroom applications (S_GENCLASS)	222
12.12.Item wording for the scale Specialist classroom applications (S_SPECLASS)	223
12.13.Item parameters, item fits, and WRMSD for the scale Specialist classroom applications (S_SPECLASS)	223
12.14.Item wording for the scale General ICT self-efficacy (S_GENEFF)	224
12.15.Item parameters, item fits, and WRMSD for the scale General ICT self-efficacy (S_GENEFF)	224
12.16.Item wording for the scale Specialist ICT self-efficacy (S_SPECEFF)	225
12.17.Item parameters, item fits, and WRMSD for the scale Specialist ICT self-efficacy (S_SPECEFF)	225
12.18.Item wording for the scale Attitudes toward ICT - positive outcomes general (S_ICTPOSG)	226
12.19.Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - positive outcomes general (S_ICTPOSG)	226
12.20.Item wording for the scale Attitudes toward ICT - negative outcomes (S_ICTNEG)	226
12.21.Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - negative outcomes (S_ICTNEG)	227
12.22.Item wording for the scale Attitudes toward ICT - expectations for future use of ICT (S_ICTFUT)	227
12.23.Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - expectations for future use of ICT (S_ICTFUT)	227
12.24.Item wording for the scale Attitudes toward ICT - positive outcomes (S_ICTPOSS)	228
12.25.Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - positive outcomes (S_ICTPOSS)	228
12.26.Item wording for the scale Learning about CT in class (S_CTCLS)	229
12.27.Item parameters, item fits, and WRMSD for the scale Learning about CT in class (S_CTCLS)	229
12.28.Item wording for the scale Teachers' ICT self-efficacy (T_ICTEFF)	230
12.29.Item parameters, item fits, and WRMSD for the scale Teachers' ICT self-efficacy (T_ICTEFF)	230
12.30.Item wording for the scale Emphasis on ICT use in initial teacher education (T_EMPITE)	231
12.31.Item parameters, item fits, and WRMSD for the scale Emphasis on ICT use in initial teacher education (T_EMPITE)	231
12.32.Item wording for the scale Teacher participation in professional learning activities (T_PROFLRN)	232
12.33.Item parameters, item fits, and WRMSD for the scale Teacher participation in professional learning activities (T_PROFLRN)	232
12.34.Item wording for the scale Teacher perceptions of need for professional learning (T_PROFNEED)	233
12.35.Item parameters, item fits, and WRMSD for the scale Teacher perceptions of need for professional learning (T_PROFNEED)	233
12.36.Item wording for the scale Collaboration between teachers in using ICT (T_COLICT)	234
12.37.Item parameters, item fits, and WRMSD for the scale Collaboration between teachers in using ICT (T_COLICT)	234
12.38.Item wording for the scale Availability of computer resources at school (T_RESRC)	235
12.39.Item parameters, item fits, and WRMSD for the scale Availability of computer resources at school (T_RESRC)	235
12.40.Item wording for the scale Negative views on using ICT in teaching and learning (T_VWNEG)	236

12.41.Item parameters, item fits, and WRMSD for the scale Negative views on using ICT in teaching and learning (T_VWNEG)	236
12.42.Item wording for the scale Positive views on using ICT in teaching and learning (T_VWPOS)	237
12.43.Item parameters, item fits, and WRMSD for the scale Positive views on using ICT in teaching and learning (T_VWPOS)	237
12.44.Item wording for the scale Teacher centered teaching approach (T_TEAPTC)	238
12.45.Item parameters, item fits, and WRMSD for the scale Teacher centered teaching approach (T_TEAPTC)	238
12.46.Item wording for the scale Inquiry-based teaching approach (T_TEAPIN)	239
12.47.Item parameters, item fits, and WRMSD for the scale Inquiry-based teaching approach (T_TEAPIN)	239
12.48.Item wording for the scale Experiential/scenario-based teaching approach (T_TEAPEX)	240
12.49.Item parameters, item fits, and WRMSD for the scale Experiential/scenario-based teaching approach (T_TEAPEX)	240
12.50.Item wording for the scale Teachers' use of ICT for classroom activities - basic tasks (T_ICTCLASB)	240
12.51.Item parameters, item fits, and WRMSD for the scale Teachers' use of ICT for classroom activities - basic tasks (T_ICTCLASB)	241
12.52.Item wording for the scale Teachers' use of ICT for classroom activities - advanced tasks (T_ICTCLASA)	241
12.53.Item parameters, item fits, and WRMSD for the scale Teachers' use of ICT for classroom activities - advanced tasks (T_ICTCLASA)	242
12.54.Item wording for the scale Use of digital learning tools (T_USETOOL)	242
12.55.Item parameters, item fits, and WRMSD for the scale Use of digital learning tools (T_USETOOL)	243
12.56.Item wording for the scale Use of general utility software (T_USEUTIL)	243
12.57.Item parameters, item fits, and WRMSD for the scale Use of general utility software (T_USEUTIL)	244
12.58.Item wording for the scale Emphasis on ICT capabilities in class (T_ICTEMP)	244
12.59.Item parameters, item fits, and WRMSD for the scale Emphasis on ICT capabilities in class (T_ICTEMP)	245
12.60.Item wording for the scale Teacher emphasis on CT related tasks in class (T_CODEMP)	245
12.61.Item parameters, item fits, and WRMSD for the scale Teacher emphasis on CT related tasks in class (T_CODEMP)	246
12.62.Item wording for the scale Teacher epistemological beliefs-Embodied cognition (T_EPIEMC)	246
12.63.Item parameters, item fits, and WRMSD for the scale Teacher epistemological beliefs-Embodied cognition (T_EPIEMC)	247
12.64.Item wording for the scale Teacher epistemological beliefs-Cognitivism (T_EPICOG)	247
12.65.Item parameters, item fits, and WRMSD for the scale Teacher epistemological beliefs-Cognitivism (T_EPICOG)	248
12.66.Item wording for the scale Teacher epistemological beliefs-Constructivist (T_EPICON)	248
12.67.Item parameters, item fits, and WRMSD for the scale Teacher epistemological beliefs-Constructivist (T_EPICON)	248
12.68.Item wording for the scale Principals' views on using ICT for educational outcomes (P_VWICT)	249
12.69.Item parameters, item fits, and WRMSD for the scale Principals' views on using ICT for educational outcomes (P_VWICT)	249
12.70.Item wording for the scale Principals' reports on expectations of ICT use by teachers (P_EXPLRN)	250

12.71.Item parameters, item fits, and WRMSD for the scale Principals' reports on expectations of ICT use by teachers (P_EXPLRN)	250
12.72.Item wording for the scale Principals' reports on expectations for teacher collaboration using ICT (P_EXPTCH)	251
12.73.Item parameters, item fits, and WRMSD for the scale Principals' reports on expectations for teacher collaboration using ICT (P_EXPTCH)	251
12.74.Item wording for the scale Principals' reports on priorities for facilitating use of ICT (P_PRIORS)	252
12.75.Item parameters, item fits, and WRMSD for the scale Principals' reports on priorities for facilitating use of ICT (P_PRIORS)	252
12.76.Item wording for the scale Principals' use of ICT for general school-related activities (P_ICTUSE)	253
12.77.Item parameters, item fits, and WRMSD for the scale Principals' use of ICT for general school-related activities (P_ICTUSE)	253
12.78.Item wording for the scale Principals' use of ICT for school-related communication activities (P_ICTCOM)	254
12.79.Item parameters, item fits, and WRMSD for the scale Principals' use of ICT for school-related communication activities (P_ICTCOM)	254
12.80.Item wording for the scale Principals' views on negative impact of ChatGPT on students' learning (P_CIMNEG)	255
12.81.Item parameters, item fits, and WRMSD for the scale Principals' views on negative impact of ChatGPT on students' learning (P_CIMNEG)	255
12.82.Item wording for the scale Principals' views on positive impact of ChatGPT on students' learning (P_CIMPOS)	256
12.83.Item parameters, item fits, and WRMSD for the scale Principals' views on positive impact of ChatGPT on students' learning (P_CIMPOS)	256
12.84.Item wording for the scale Principals' views on increased workload as a consequence of ChatGPT (P_CWORK)	257
12.85.Item parameters, item fits, and WRMSD for the scale Principals' views on increased workload as a consequence of ChatGPT (P_CWORK)	257
12.86.Item wording for the scale Principals' views on negative consequences of ChatGPT on teacher work (P_CONEG)	258
12.87.Item parameters, item fits, and WRMSD for the scale Principals' views on negative consequences of ChatGPT on teacher work (P_CONEG)	258
12.88.Item wording for the scale Principals' views on positive consequences of ChatGPT on teacher work (P_COPOS)	259
12.89.Item parameters, item fits, and WRMSD for the scale Principals' views on positive consequences of ChatGPT on teacher work (P_COPOS)	259
12.90.Item wording for the scale ICT coordinators' reports on availability of ICT resources at school (C_ICTRES)	260
12.91.Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on availability of ICT resources at school (C_ICTRES)	260
12.92.Item wording for the scale ICT coordinators' reports on availability of ICT resources at school - tools (C_RESTOOL)	261
12.93.Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on availability of ICT resources at school - tools (C_RESTOOL)	261
12.94.Item wording for the scale ICT coordinators' reports on availability of ICT resources at school - utility software (C_RESUTIL)	262
12.95.Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on availability of ICT resources at school - utility software (C_RESUTIL)	262

12.96.Item wording for the scale ICT coordinators' reports on computer resource hindrances to the use of ICT in teaching and learning (C_HINRES)	263
12.97.Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on computer resource hindrances to the use of ICT in teaching and learning (C_HINRES)	263
12.98.Item wording for the scale ICT coordinators' reports on pedagogical resource hindrances to the use of ICT in teaching and learning (C_HINPED)	264
12.99.Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on pedagogical resource hindrances to the use of ICT in teaching and learning (C_HINPED)	264
13.1. Number of jackknife zones in national samples	269
13.2. Example of computation of jackknife replicate weights	270
13.3. CIL mean variance components and reported standard error per country	272
13.4. CT variance components and reported standard error by country	273
13.5. Equating error for CIL and CT achievement estimates between cycles	277
13.6. Estimates of equating error for percentages of students at or above CIL Level 2	278
B.1. Allocation of student sample in Austria	289
B.2. Allocation of teacher sample in Austria	290
B.3. Allocation of student sample in Azerbaijan	291
B.4. Allocation of teacher sample in Azerbaijan	291
B.5. Allocation of student sample in Belgium (Flemish)	292
B.6. Allocation of teacher sample in Belgium (Flemish)	292
B.7. Allocation of student sample in Bosnia and Herzegovina	293
B.8. Allocation of teacher sample in Bosnia and Herzegovina	293
B.9. Allocation of student sample in Chile	294
B.10. Allocation of teacher sample in Chile	294
B.11. Allocation of student sample in Chinese Taipei	295
B.12. Allocation of teacher sample in Chinese Taipei	295
B.13. Allocation of student sample in Croatia	296
B.14. Allocation of teacher sample in Croatia	296
B.15. Allocation of student sample in Cyprus	297
B.16. Allocation of teacher sample in Cyprus	298
B.17. Allocation of student sample in Czech Republic	299
B.18. Allocation of teacher sample in Czech Republic	300
B.19. Allocation of student sample in Denmark	301
B.20. Allocation of teacher sample in Denmark	301
B.21. Allocation of student sample in Finland	302
B.22. Allocation of teacher sample in Finland	303
B.23. Allocation of student sample in France	304
B.24. Allocation of teacher sample in France	304
B.25. Allocation of student sample in Germany	305
B.26. Allocation of teacher sample in Germany	305
B.27. Allocation of student sample in Greece	306
B.28. Allocation of teacher sample in Greece	307
B.29. Allocation of student sample in Hungary	308
B.30. Allocation of teacher sample in Hungary	308
B.31. Allocation of student sample in Italy	309
B.32. Allocation of teacher sample in Italy	309
B.33. Allocation of student sample in Kazakhstan	310
B.34. Allocation of teacher sample in Kazakhstan	311
B.35. Allocation of student sample in Republic of Korea	312

B.36. Allocation of teacher sample in Republic of Korea	312
B.37. Allocation of student sample in Kosovo	313
B.38. Allocation of teacher sample in Kosovo	313
B.39. Allocation of student sample in Latvia	314
B.40. Allocation of teacher sample in Latvia	314
B.41. Allocation of student sample in Luxembourg	315
B.42. Allocation of teacher sample in Luxembourg	315
B.43. Allocation of student sample in Malta	316
B.44. Allocation of teacher sample in Malta	316
B.45. Allocation of student sample in the Netherlands	317
B.46. Allocation of teacher sample in the Netherlands	317
B.47. Allocation of student sample in Norway	318
B.48. Allocation of teacher sample in Norway	318
B.49. Allocation of student sample in Oman	319
B.50. Allocation of teacher sample in Oman	320
B.51. Allocation of student sample in Portugal	321
B.52. Allocation of teacher sample in Portugal	321
B.53. Allocation of student sample in Romania	322
B.54. Allocation of teacher sample in Romania	322
B.55. Allocation of student sample in Serbia	323
B.56. Allocation of teacher sample in Serbia	324
B.57. Allocation of student sample in Slovak Republic	325
B.58. Allocation of teacher sample in Slovak Republic	325
B.59. Allocation of student sample in Slovenia	326
B.60. Allocation of teacher sample in Slovenia	326
B.61. Allocation of student sample in Spain	327
B.62. Allocation of teacher sample in Spain	327
B.63. Allocation of student sample in Sweden	328
B.64. Allocation of teacher sample in Sweden	329
B.65. Allocation of student sample in United States	330
B.66. Allocation of teacher sample in United States	330
B.67. Allocation of student sample in Uruguay	331
B.68. Allocation of teacher sample in Uruguay	331
B.69. Allocation of student sample in North-Rhine Westphalia (Germany)	332
B.70. Allocation of teacher sample in North-Rhine Westphalia (Germany)	332
C.1. Scoring information for CIL module Breathing	333
C.2. Scoring information for CIL module School trip	342
D.1. Scoring information for CIL module Automated bus	349
D.2. Example scoring of the correctness of student coding responses	354
D.3. Example scoring of the efficiency of student coding responses	355
D.4. Example scoring of correctness and efficiency combined	356
D.5. Scoring information for CIL module Farm drone	356
E.1. Background variables used for conditioning	362
F.1. National item adjudication	368
G.1. Recoding rules for CIL module Breathing	369
G.2. Scoring and recoding rules for CIL module School trip	369
G.3. Scoring rubric for CT module Automated bus	370
G.4. Scoring rubric for CT module Farm drone	371

H.1. Reliabilities by country for scales from student questionnaire (part 1)	375
H.2. Reliabilities by country for scales from student questionnaire (part 2)	376
H.3. Reliabilities by country for scales from student questionnaire (part 3)	377
H.4. Reliabilities by country for scales from teacher questionnaire (part 1)	378
H.5. Reliabilities by country for scales from teacher questionnaire (part 2)	379
H.6. Reliabilities by country for scales from teacher questionnaire (part 3)	380
H.7. Reliabilities by country for scales from teacher questionnaire (part 4)	381
H.8. Reliabilities for the pooled international dataset for principal and ICT coordinator scales	382
H.9. Proportion of explained variance by the first principal component by country for scales from student questionnaire (part 1)	383
H.10. Proportion of explained variance by the first principal component by country for scales from student questionnaire (part 2)	384
H.11. Proportion of explained variance by the first principal component by country for scales from student questionnaire (part 3)	385
H.12. Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 1)	386
H.13. Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 2)	387
H.14. Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 3)	388
H.15. Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 4)	389
H.16. Proportion of explained variance by the first principal component for the pooled international dataset for principal and ICT coordinator scales	390

Chapter 1:

Introduction

Julian Fraillon

1.1 Background

The International Association for the Evaluation of Educational Achievement's (IEA's) International Computer and Information Literacy Study (ICILS) was established in response to the recognition of the essential role that information and communication technology (ICT) has come to play in modern society and the consequent need for young people to develop relevant capabilities to participate safely, productively, and effectively in the digital world. The first cycle of ICILS was conducted in 2013, with the second cycle in 2018. ICILS 2023 builds on the work of these two ICILS cycles.

Since 2013, ICILS has measured the achievement of grade 8 (or equivalent) students in computer and information literacy (CIL) which is defined as an "individual's ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in society" (Fraillon & Duckworth, [2024](#), p. 26).

Since 2018, participating countries could also choose for students to complete an optional additional assessment of computational thinking (CT) which is defined as an "individual's ability to recognize aspects of real-world problems that are appropriate for computational formulation and to evaluate and develop algorithmic solutions to those problems so that the solutions could be operationalized with a computer" (Duckworth & Fraillon, [2024](#), p. 38).

In ICILS 2023, students in 34 countries and one benchmarking participant completed the CIL assessment, and students from 24 countries and one benchmarking participant completed the additional CT assessment.

In addition to measuring variations in CIL and CT among and within countries, ICILS collects and reports on extensive data representing the contexts in which student digital literacy capabilities are being developed within and outside of school. These contextual data are collected from national study centers, school principals, school ICT coordinators, teachers, and students in schools participating in the study.

The ICILS international database (IDB) is publicly available for each of the ICILS study cycles, so that ICILS data may be further analyzed and reported to contribute to our understanding of digital literacy education and to inform policy and practice. The ICILS IDBs can be accessed from [IEA's data repository](#).

This report aims to provide comprehensive details of the implementation of ICILS 2023 including instrument development and preparation processes, data collection procedures (including sampling and field operations), and data management, analysis, and reporting procedures. Information in this report supplements and is supplemented by information provided in:

- The [ICILS 2023 assessment framework](#)
- The [ICILS 2023 international report](#)
- The [ICILS 2023 IDB](#) and [user guide](#)

Some of the processes and procedures described in the chapters of this report are very similar to, or even identical to those that were used in previous cycles of ICILS (see Fraillon et al., [2020](#); Fraillon et al., [2014](#)). Where this is the case, the descriptions of these processes and procedures from previous



ICILS technical reports were used as the basis for the content in this report. These descriptions were reviewed and, where necessary, revised. However, where it was deemed appropriate to do so, the text from previous technical reports may have been included in this report without change. This text is not cited, as it has been judged by the ICILS International Study Center (ISC), including the editors of this report, to be an accurate description of what was completed in ICILS 2023.

1.2 Study overview

Study coordination and administration

International Study Center

ICILS 2023 was a large cross-national research project in which a suite of instruments were first centrally developed to provide comparable data across countries, and then administered to representative samples of students, teachers, ICT-coordinators, and school principals within participating countries. The corresponding data were then cleaned, coded, scored, analyzed, and reported.

The study administration was correspondingly established and managed with close coordination at the international and national levels. The ICILS ISC was located at IEA, and national study centers were located in each participating country.

The ISC comprises IEA staff, led by the ICILS International Study Director, with responsibility for the development of the international instrument content used in the study, as well as oversight of the operational procedures to be conducted within participating countries. The role of the ISC includes, but is not limited to:

- Conceptual leadership of the study, including development of the ICILS assessment framework and the ICILS international test and questionnaire instruments
- Operational leadership of the study, including international study coordination, and the development and oversight of operational procedures in countries including:
 - Sampling of schools within countries and of participants within schools
 - Instrument adaptation and translation
 - Field operations including data collection within countries
 - Quality assurance of the implementation of operational procedures within countries
 - Scoring and coding selected student responses to test and questionnaire items
 - Data management and adjudication
- Communication with, and support for participating countries, including development of international manuals and support materials, and provision of training
- Scaling and data analysis
- Dissemination of ICILS information, including reporting international results, their consequences, and informing policy and practice with respect to the findings of ICILS

As ICILS is an IEA study, the study operations are based on and consistent with established IEA policies and protocols. However, the ISC has the capacity, where appropriate, to tailor operational procedures and practices to suit the specific needs of the study. The technical implementation of the study, including its adherence to IEA policies, protocols, and technical standards is overseen by the IEA Technical Executive Group (TEG).

The ISC also operates in collaboration with external experts as required. Formally, one body, the Questionnaire Development Group (QDG), was established to support development of the ICILS questionnaire instruments (see [Chapter 4](#)).

**National study centers**

Within each participating country, the national study center (hereafter referred to as *national center*) is responsible for the coordination and implementation of the study. National centers are appointed within each country according to their internal policies and procedures. This process is independent of IEA and the ICILS ISC. The work of the national center in each country is led by the country-appointed national research coordinator¹ (NRC). The NRC is the primary point of contact with the ISC and is expected to manage ICILS-related activities within the country.

National centers, under the leadership of NRCs and with the guidance and support of the ISC, complete the sampling, instrument preparation, and data collection activities listed above. NRCs, or their delegates, are also invited to contribute to the development of the ICILS assessment framework and instrument content in collaboration with the ISC. Many national centers also disseminate ICILS results with a focus on their own national findings by authoring national reports and providing associated advice relating to policy and practice in their own national contexts.

ICILS assessment framework

The contents and high-level operational procedures of ICILS 2023 are explained and described in the ICILS 2023 assessment framework (Fraillon & Rožman, 2024).

The framework includes:

- A detailed background and rationale for the study;
- The structure and content of the two ICILS achievement variables (CIL and CT);
- An explanation of the conceptual model underpinning the analytic relationship between contextual data collected in ICILS and student achievement;
- Details of the contextual data collected in ICILS; and
- Details of the student achievement test and all questionnaire instruments used in ICILS.

ICILS instruments

All ICILS instruments were delivered on computer. Three different delivery platforms were used, one platform was used for the student test and questionnaire instruments; a second platform was used to deliver the questionnaire instruments completed by staff in sampled schools; and a third platform was used to deliver the questionnaire completed by staff in national centers (see [Chapter 3](#) for more details).

All ICILS instruments (excluding the national contexts survey—see below) were adapted and translated into the target languages used in each participating country, with the express intention that the meaning of each task was equivalent across all countries and languages. Each ICILS questionnaire (except the national contexts survey) also included a small number of questions that were optional for national centers to include or exclude in their questionnaires. National centers also had the opportunity to include a small number of additional questions of particular national interest to their questionnaire instruments. National centers were not permitted to add or remove tasks or task content from the test instrument or instruments (CIL only or CIL and CT) used in their country. All decisions about adaptations of student instrument content were completed in consultation with the ICILS ISC (see [Chapter 5](#) for more details about instrument adaptation and translation).

CIL test instrument

The CIL test instrument comprised seven 30-minute test modules. Each student completed two modules that were delivered in a fully balanced randomized design (see [Chapter 2](#) for more details). Three

¹ Countries may also appoint more than one person to the position of NRC.



of the modules were newly developed for inclusion in ICILS 2023, two modules were first used in ICILS 2018 and two modules were first used in ICILS 2013 (and again used in ICILS 2018).

CT test instrument

The CT test instrument comprised four 25-minute test modules. In countries where the CT international option was administered, each student completed two modules that were delivered in a fully balanced randomized design (see [Chapter 2](#) for more details). Two of the modules were newly developed for inclusion in ICILS 2023 and two modules were first used in ICILS 2018.

International student questionnaire

All students in the sampled classes completed a 30-minute international questionnaire.

International school staff questionnaires

Three separate questionnaires were developed to be completed by staff in sampled ICILS schools:

- A 30-minute teacher questionnaire: This was designed to be completed by a random sample of 15 teachers of grade 8 students in each sampled school.
- A 20-minute ICT coordinator questionnaire: This was designed to be completed by the designated ICT coordinator in each sampled school.
- A 20-minute principal questionnaire.

National contexts survey

ICILS national research coordinators (NRCs) oversaw completion of the national contexts survey (NCS) within each participating country. This questionnaire included detailed information regarding the education system, plans, and policies with respect to CIL and CT education within each country. It was expected that NRCs would seek advice from relevant national experts or invite suitably qualified personnel to provide answers for the questions in the NCS. The NCS was presented in English only and any text responses were required to be provided in English.

ICILS participants

Education systems participating in 2023

The following 34 countries and one benchmarking participant took part in ICILS 2023.



Table 1.1: ICILS 2023 participating education systems

Countries		
Austria (CIL&CT)	Germany (CIL&CT)	Oman (CIL&CT)
Azerbaijan (CIL)	Greece (CIL)	Portugal (CIL&CT)
Belgium (Flemish) (CIL&CT)	Hungary (CIL)	Romania (CIL)
Bosnia and Herzegovina (CIL)	Italy (CIL&CT)	Serbia (CIL&CT)
Chile (CIL)	Kazakhstan (CIL)	Slovak Republic (CIL&CT)
Chinese Taipei (CIL&CT)	Korea (Rep. of) (CIL&CT)	Slovenia (CIL&CT)
Croatia (CIL&CT)	Kosovo (CIL)	Spain (CIL)
Cyprus (CIL)	Latvia (CIL&CT)	Sweden (CIL&CT)
Czech Republic (CIL&CT)	Luxembourg (CIL&CT)	United States (CIL&CT)
Denmark (CIL&CT)	Malta (CIL&CT)	Uruguay (CIL&CT)
Finland (CIL&CT)	Netherlands (CIL&CT)	
France (CIL&CT)	Norway (CIL&CT)	
Benchmarking participant		
North Rhine-Westphalia (Germany) (CIL&CT)		

Note

(CIL): the education system administered the core CIL assessment only.

(CIL&CT): the education system administered both the core CIL and optional CT assessment.

Education systems participating in both 2023 and in previous cycles

Two countries, Germany and Korea (Rep. of), participated in all three cycles of ICILS, and met the necessary sample participation requirements within each cycle. In these two countries, valid comparisons of students' CIL achievement can be made across the three ICILS cycles.

Eight countries, Denmark, Finland, France, Italy, Kazakhstan, Luxembourg, Portugal, Uruguay, and the benchmarking participant North Rhine-Westphalia (Germany) participated in ICILS 2018 and 2023 and met the sample participation requirements within each cycle. In these eight countries and the benchmarking participant, valid comparisons of students' CIL achievement can be made between the 2018 and 2023 cycles.

Valid comparisons of student CT achievement between the 2018 and 2023 cycles can also be made in all of these participants except Uruguay, where CT was not assessed in 2018.

A further five countries, Croatia, the Czech Republic, Norway, the Slovak Republic, and Slovenia participated in ICILS 2013 and 2023 and met the sample participation requirements within each cycle. In these five countries, valid comparisons of students' CIL achievement can be made between ICILS 2013 and ICILS 2023.

Population definitions and sampling

The samples were designed as two-stage cluster samples. During the first stage of sampling, schools with enrolled students at the target grade were randomly selected with a probability proportional to size as measured by the number of students enrolled in a school. The numbers required in the sample to achieve the necessary precision were estimated on the basis of national characteristics.

The ICILS student population was defined as students in grade 8 (typically about 14 years of age in most countries), provided that the average age of students in this grade was at least 13.5 years at the time of the assessment. Grade level equivalence and student age are shown by country in [Table 1.2](#).



Table 1.2: Grade level equivalence and student age across countries

Country	Grade level equivalence	Average age of students in years
¹ Austria	Grade 8	14.5
Azerbaijan	Grade 8	14.3
[†] Belgium (Flemish)	Grade 8	14.0
³ Bosnia and Herzegovina	Grade 8	14.2
Chinese Taipei	Grade 8	14.3
¹ Croatia	Grade 8	14.7
Cyprus	Grade 8	13.9
¹ Czech Republic	Grade 8	14.3
^{†1} Denmark	Grade 8	14.9
Finland	Grade 8	14.8
France	Grade 8	13.9
Germany	Grade 8	14.6
Greece	Grade 8	13.9
Hungary	Grade 8	14.6
Italy	Grade 8	13.8
¹ Kazakhstan	Grade 8	14.3
[†] Korea, Republic of	Grade 8	14.4
¹ Kosovo	Grade 8	14.1
¹ Latvia	Grade 8	14.9
Luxembourg	Grade 8	14.4
Malta	Grade 8	13.9
¹ Norway (Grade 9)	Grade 9	14.8
Oman	Grade 8	13.8
¹ Portugal	Grade 8	14.1
^{†12} Romania	Grade 8	14.4
¹ Serbia	Grade 8	14.7
Slovak Republic	Grade 8	14.3
¹ Slovenia	Grade 8	14.0
¹ Spain	Grade 8	14.0
¹ Sweden	Grade 8	14.8
[†] Uruguay	Grade 8	14.1
Benchmarking participant		
¹ North Rhine-W. (Germany)	Grade 8	14.5
Countries not meeting sample participation requirements		
[§] Netherlands	Grade 8	14.2
[†] United States	Grade 8	14.3

Notes:

[†] Met guidelines for sampling participation rates only after replacement schools were included.[†] Did not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.[§] Weighted participation rate was less than 50%; as these data do not meet IEA reporting standards, we urge great caution when interpreting the results and strongly discourage comparison of these data with data from other participating countries.¹ National defined population covers 90% to 95% of the national target population.² Country surveyed target grade in the first half of the school year.³ National defined population covers 61% of the national target population.

Within each school, one class was randomly sampled from all the classes at the target grade. All students within each sampled class were then invited to participate.

The ICILS teacher population was defined as consisting of all teachers teaching regular school subjects to the students in the target grade at each sampled school. It included only those teachers who were teaching the target grade during the testing period and who had been employed at the school since



the beginning of the school year.

In schools with more than 20 teachers at the target grade (grade 8 or equivalent), 15 teachers of these teachers were to be selected at random to participate. In schools with 20 or fewer teachers at the target grade, all teachers teaching the target grade were invited to participate.

The principal and the nominated ICT coordinator (either by title or designated by the school principal) in each sampled school also were invited to participate.

ICILS 2023 collected data from 132,889 grade 8 (or equivalent) students from 5,299 schools across the 34 participating countries and one benchmarking participant. 132,652 students completed the CIL² assessment and 85,240 completed the CT assessment. Data were collected from 60,835 teachers in the sampled schools.

Complete details of the population definitions, sampling and participation are provided in [Chapter 6](#) and [Chapter 7](#).

Administration periods

The ICILS 2023 main survey data collection took place in the first half of 2023 in countries using a Northern Hemisphere school calendar³ and in the second half of 2023 for participants using a Southern Hemisphere school calendar.

1.3 Outline of this technical report

This introductory chapter is followed by 12 chapters. Chapters 2, 3, and 4, address the development and content of the instruments and software systems that were used in the study. Chapter 2 focuses on the tests of CIL and CT. Chapter 3 provides a high-level description of the software systems used across ICILS. Chapter 4 details the development of the questionnaires used in ICILS for gathering data from students, teachers, principals, and school ICT coordinators. The chapter also provides an outline of the development of the national contexts survey, completed by the national research coordinators.

Chapters 5 through 9 focus on the implementation of the survey in ICILS 2023. Chapter 5 describes the translation procedures and national adaptations used in ICILS. Chapter 6 details the sampling design and implementation processes. The outcomes of these processes, including the establishment of sampling weights and the achieved participation rates are described in Chapter 7. Chapter 8 describes the field operations procedures that underpinned data collection in ICILS 2023, and Chapter 9 describes the quality control procedures used to collect information and report on the success of the implementation of field operations procedures within and across countries.

Chapters 10, 11, 12, and 13 address data management, analysis and reporting. Chapter 10 describes the data-management processes that resulted in the creation of the ICILS IDB. Chapter 11 details the scaling procedures for the tests of CIL and CT and Chapter 12 describes the scaling procedures for the questionnaire items (student, teacher, principal, and ICT coordinator questionnaires). The final chapter, Chapter 13, presents an account of the analyses that underpinned the results reported in the ICILS 2023 international report (Fraillon, 2025).

References

Duckworth, D., & Fraillon, J. (2024). Computational thinking framework. In J. Fraillon & M. Rožman (Eds.), *IEA International Computer and Information Literacy Study 2023 assessment framework* (pp. 35–43). Springer. https://doi.org/10.1007/978-3-031-61194-0_3

² Slightly fewer students completed the core CIL assessment than provided data because a small number of students completed the CT assessment without having completed the CIL assessment.

³ Except Romania where data collection took place in the second half of 2023.



- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (Eds.). (2020). *IEA International Computer and Information Literacy Study 2018 technical report*. International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>
- Fraillon, J. (Ed.). (2025). *An international perspective on digital literacy: Results from ICILS 2023*. Springer. <https://doi.org/10.1007/978-3-031-87722-3>
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA International Computer and Information Literacy Study international report*. Springer Nature. <https://doi.org/10.1007/978-3-319-14222-7>
- Fraillon, J., & Duckworth, D. (2024). Computer and information literacy framework. In J. Fraillon & M. Rožman (Eds.), *IEA International Computer and Information Literacy Study 2023 assessment framework* (pp. 21–34). Springer. https://doi.org/10.1007/978-3-031-61194-0_2
- Fraillon, J., & Rožman, M. (2024). *IEA International Computer and Information Literacy Study 2023 assessment framework*. Springer. <https://doi.org/10.1007/978-3-031-61194-0>

Chapter 2:

ICILS 2023 test development

Daniel Duckworth and Julian Fraillon

2.1 Introduction

This chapter provides a detailed description of the test development process and review procedures, as well as the test design implemented for the ICILS 2023 field trial and main survey. The processes are relevant for the tests of both computer and information literacy (CIL) and computational thinking (CT) and are consequently described together. Where relevant, details of the content of each test are described separately.

The development of new content for the ICILS 2023 test instruments occurred over a 24-month period, from January 2020 to December 2022. The majority of this work was conducted by the International Study Center (ISC) at IEA, in collaboration with National Research Coordinators (NRCs) and other research partners. All participating students completed the test of CIL. Additionally, the test of CT was administered to students in countries that chose to participate in this optional assessment.

2.2 Terminology

This chapter refers to key terms related to the ICILS assessments of CIL and CT. The definitions and relationships between these terms are provided below, with examples to illustrate their application.

- **Modules:** A module is a set of assessment tasks centered around an authentic theme, structured in a linear narrative. For example, in the module Breathing, students managed files and collected and evaluated information to create a presentation explaining the process of breathing to eight- or nine-year-old students.
- **Tasks:** A task refers to an individual assessment activity presented to students. Each task includes stimulus material, instructions, and may include actions required to achieve an objective or a question to answer. For example, in task two of the Breathing module, the instructions directed students to save a new presentation as 'Breathing,' which required them to use the 'Save as' function (see Duckworth, 2025a, p. 288).
- **Scoring variables:** A scoring variable captures the codified student response to a task. The structure of scoring variables varies by task type:
 - Simple tasks (Duckworth & Fraillon, 2024b, p. 62) typically use a single scoring variable per task, indicating whether the response is correct or incorrect.
 - Complex tasks (Duckworth & Fraillon, 2024b, p. 66) often require multiple scoring variables. For example, in the Breathing authoring task where students create a presentation, scoring variables measure content accuracy, design quality, and organization separately (see Appendix C).
 - Composite variables: Some scoring variables are derived from multiple component variables. For instance, in task four of the module School trip (Duckworth, 2025a, p. 262), students were required to enter a specific value into a specific cell in a spreadsheet. One component variable recorded whether the correct value was entered into any cell. A second component variable recorded whether any value was entered into the correct cell (see Appendix C). These details were combined into a composite variable representing a single score for the task (see Chapter 11). A full credit score of 2 was assigned if both component variables reflected the completed actions, a partial credit score of 1 was assigned if



either one of the component variables reflected the completed actions, and a score of 0 was assigned if neither component variable reflected the completed actions.

- Score points: Score points define the possible values assigned to a scoring variable. The number of score points varies by task type:
 - Binary scoring (e.g., correct = 1, incorrect = 0) applies to simple tasks.
 - Partial credit scoring applies to more complex tasks. For example, one of the scoring variables associated with an authoring task may assign a score of 2 for a full credit response, 1 for a partial credit response, and 0 for an incorrect response.
- Items: An item represents a unit of measurement derived from scored student responses. Items serve as the basis for constructing the CIL and CT achievement scales (see [Chapter 11](#)).

2.3 Test scope and format

ICILS assessment framework

The ICILS student tests of CIL and CT were developed with reference to the ICILS assessment framework.⁴ According to the framework, CIL is defined as an “individual’s ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in society” (Fraillon & Duckworth, 2024, p. 26). CT is defined as the “ability to recognize aspects of real-world problems which are appropriate for computational formulation and to evaluate and develop algorithmic solutions to those problems so that the solutions could be operationalized with a computer” (Duckworth & Fraillon, 2024a, p. 38).

The framework describes CIL and CT as comprising strands, each of which is specified in terms of several aspects. A strand refers to the overarching conceptual category for framing the skills and knowledge addressed by the test instruments. An aspect refers to the specific content category within a strand. This organizational structure was established to support users to access and understand breadth and scope of the content of each construct. Establishment of the strands and aspects in each of CIL and CT was based on the substantive content of each domain and did not presuppose that either construct would comprise more than one measurement dimension. In ICILS 2013 and ICILS 2018, multidimensional IRT modeling was used to examine the potential dimensionality of CIL, with ICILS 2018 also assessing the dimensionality of CT. In ICILS 2023, the dimensionality of CIL was re-evaluated, and additional analyses were conducted for CT (see [Chapter 11](#)). In each case the latent correlations between potential sub-dimensions were very high and each of CIL and CT were reported in ICILS as unidimensional achievement constructs (Fraillon et al., 2015, 2020).

The following lists outline the strands and corresponding aspects of the CIL and CT constructs. Full details of the constructs can be found in the ICILS 2023 assessment framework (Fraillon & Rožman, 2024).

The CIL construct

- Strand 1: Understanding computer use, comprising two aspects:
 - Aspect 1.1: Foundations of computer use
 - Aspect 1.2: Computer use conventions
- Strand 2: Gathering information, comprising two aspects:
 - Aspect 2.1: Accessing and evaluating information
 - Aspect 2.2: Managing information
- Strand 3: Producing information, comprising two aspects:

4 The framework can be downloaded from: <https://www.springer.com/gp/book/9783030193881>



- Aspect 3.1: Transforming information
- Aspect 3.2: Creating information
- Strand 4: Digital communication, comprising two aspects:
 - Aspect 4.1: Sharing information
 - Aspect 4.2: Using information responsibly and safely

The CT construct

- Strand 1: Conceptualizing problems, comprising three aspects:
 - Aspect 1.1: Knowing about and understanding digital systems
 - Aspect 1.2: Formulating and analyzing problems
 - Aspect 1.3: Collecting and representing relevant data
- Strand 2: Operationalizing solutions, comprising two aspects:
 - Aspect 2.1: Planning and evaluating solutions
 - Aspect 2.2: Developing algorithms, programs, and interfaces

The ICILS test instruments

The CIL test instrument

The ICILS main survey CIL test instrument comprised seven modules, each designed to be completed in 30 minutes. Each module has a series of small self-contained tasks, designed to be able to be completed quickly (usually in less than one minute), and a large authoring task designed to be completed in 10 to 15 minutes. A module presents the small tasks as a combination of skill execution and information management activities that students must complete in preparation for the large task. Four of the modules were secure trend modules, two from ICILS 2013 and two from ICILS 2018, which provided a basis for reporting all CIL data collected in ICILS 2023 on the ICILS CIL achievement scale established in 2013. The remaining three modules were newly developed specifically for ICILS 2023. Each student completed a randomly assigned pair of CIL test modules out of the seven available modules in a fully balanced rotational design.

At the start of each module, students were presented with an overview of the theme, the purpose of the tasks, and a basic description of the large task they would complete. Students were required to complete the tasks in the given sequence and could not revisit any tasks once completed. [Table 2.1](#) provides a summary of the seven CIL test modules.



Table 2.1: Summary of ICILS 2023 main survey CIL test modules

Module title (study cycle)	Description
Board games club (2018, 2023)	Students use a school-based social network for direct messaging and group posts to encourage peers to join a board games interest group.
Breathing (2013, 2018, 2023)	Students manage files and collect and evaluate information to create a presentation explaining the process of breathing to eight-or-nine-year-old students.
Computer use and health (2023)	Students collaborate with a partner using a chat app to communicate, manage files, and access and evaluate information sources for a research report on computer use and health issues.
Internet safety (2023)	Students research information on identifying fake information and scams, personal information security, and reporting suspicious content to authorities. They work on a project to create a digital poster for younger students at their school, providing guidelines and tips for avoiding scams.
Paper books vs ebooks (2023)	Students use the internet to find people's opinions about their preferences for paper books and ebooks. They then write a set of notes comparing the positives and negatives of each, which someone else can use to prepare a presentation.
Recycling (2018, 2023)	Students access and evaluate information from a video-sharing website to identify a suitable source related to waste reduction, reuse, and recycling. They take research notes from the video and use these notes as the basis for designing an infographic to raise awareness about waste reduction, reuse, and recycling.
School trip (2013, 2018, 2023)	Students help plan a walking tour excursion using online database tools, selecting and adapting information to produce an information sheet for their peers. The information sheet includes a map with tour directions created using a map annotation tool.

The CT test instrument

The ICILS main survey CT test instrument consisted of four modules, each designed to be completed in 25 minutes. After completing the CIL test and the student questionnaire, each student completed a randomly assigned pair of CT modules out of the four modules in a fully balanced rotational design. Two of these modules were secure trend modules from ICILS 2018, while the remaining two were newly developed for ICILS 2023.

Like the CIL assessment, each CT test module was organized in a linear narrative structure and reflected a central theme. However, unlike the CIL modules, the CT modules did not culminate in a large authoring task. Instead, each module consisted of a series of tasks focused on the processes of planning and executing computer-based solutions to real-world problems. [Table 2.2](#) provides a summary



of the four CT test modules.

Table 2.2: Summary of ICILS 2023 main survey CT test modules

Module title (study cycle)	Description
Activity tracker (2023)	Students use decision trees, simulations, and block-based coding to plan, develop, evaluate, and debug components of a smartphone app that tracks a user's physical activity using data from the device's sensors.
Automated bus (2018, 2023)	Students plan various aspects of a program and configure the navigation and brake systems to operate a driverless bus using interactive directed graphs, decision trees, and simulations.
Farm drone (2018, 2023)	Students work in a block-based coding environment to create, test, and debug the conditional logic, loops, and commands that control a farming drone's actions.
Suns and moons (2023)	Students use decision trees and a block-based coding environment to create, evaluate, and debug the rules and logic of a tic-tac-toe game.

The ICILS task types

CIL task types

The ICILS CIL test modules included three broad categories of tasks described below:

- Information-based response tasks: These tasks make use of the “digital interface to deliver pencil-and-paper style questions in a slightly richer format than traditional paper-based methods” (Duckworth & Fraillon, 2024b, p. 46). The response formats for these tasks can vary (e.g., multiple choice, short constructed response, drag and drop), the stimulus material for these tasks is usually static or with minimal interactivity and the purpose of these tasks is to “capture evidence of students’ knowledge and understanding of CIL independently of students using anything beyond the most basic skills required to record a response” (Duckworth & Fraillon, 2024b, p. 62).
- Skills tasks: In these tasks students “interact with functional simulations of generic software or universal applications to accomplish a task” (Duckworth & Fraillon, 2024b, p. 66). The number of steps required to complete a skills task and the number of different correct methods for executing a task varies across skills tasks. Linear skills tasks require students to execute one or more commands in a given sequence (such as copy and paste) whereas nonlinear skills tasks require students to execute a function involving more than one sub-command without a single given sequence (such as using the filter functions in an web-based table to locate information). The linear skills tasks test students’ point-in-time knowledge of how to execute a given simple skill. Students receive feedback from the assessment system that they have completed an action, without seeing the result of their action. In linear skills tasks, students have one chance to re-complete an action (without penalty) if they wish. The nonlinear skills tasks respond to students actions as if they are operating in a live application environment. Students therefore must decide when they believe they have completed each specified non-linear skills task and prompt the system to progress to the next task. Before students progress to the next task, they have the opportunity to check and alter their response to a non-linear skills task.



- Authoring tasks: These tasks “require students to alter and create information products using authentic computer software applications” (Duckworth & Fraillon, 2024b, p. 69). The complexity of authoring tasks varies according to the number of different applications students were required to use, the degree of openness or constraint placed on the range of student responses to the task, and the amount of information students were required to evaluate and make use of when altering or creating information products. The authoring tasks were most commonly the large task within each module and typically were scored by human scorers using multiple analytic criteria.

CT task types

Like the CIL test modules, the CT test modules contain information-based response tasks and skills tasks. However, in addition to these, the CT test modules include task types that are unique to the assessment of CT.

The three CT-specific task types are described below:

- Nonlinear systems transfer tasks: These tasks require students to “interpret, transfer, and adapt algorithmic information, so that the results of applying algorithmic instructions can be visually displayed” (Duckworth & Fraillon, 2024b, p. 71). The response formats for these tasks can vary but what they have in common is that students are required to make connections between a visual representation of an algorithmic sequence and the steps of the sequence described as text.
- Simulation tasks: These tasks “require students to set parameters, run a simulation to collect data, and interpret the data to answer a research question” (Duckworth & Fraillon, 2024b, p. 72), typically as part of developing an understanding of real-world problems or to evaluate solutions to problems.
- Block-based coding tasks: These tasks require students to manipulate visual code blocks that can be used to execute a range of actions depicted in a visual display. Students could also run the code at any time and view the execution of code blocks and the corresponding visual display. They could also separately reset both the visual display and the code blocks/work space to their original states. There were two main forms of visual coding tasks:
 - Algorithm construction tasks “require students to develop their own solution to a problem by iteratively adding code blocks to the work space and executing the algorithm to see the results” (Duckworth & Fraillon, 2024b, p. 74).
 - Algorithm debugging tasks “require students to modify an existing algorithm (changing the structure of the code and parameters of the code blocks in the work space) to solve the problem presented by the task instructions” (Duckworth & Fraillon, 2024b, p. 74).

Further details of the ICILS CIL and CT test modules are presented in Chapter 5 of the ICILS 2023 assessment framework (Duckworth & Fraillon, 2024b) and detailed descriptions of tasks from released CIL and CT modules are presented in Appendix E and Appendix F of the ICILS 2023 international report (Duckworth, 2025a, 2025b).



2.4 Test-development process

The test-development process involved a series of stages that applied to both the CIL and CT test instruments. While these stages were generally sequential, the iterative and collaborative nature of the process meant that some materials were reviewed and revised multiple times within certain stages. In summary, the ISC developed item materials, occasionally incorporating suggestions from NRCs, which were then reviewed by NRCs and subsequently revised by the ISC. This process was sometimes repeated. [Table 2.3](#) provides an overview of the test development processes and timeline.

The NRC meetings and scorer training meetings were critical points in the test-development process. The first meeting of NRCs coincided with the early phases of the COVID-19 pandemic. As a consequence of the pandemic, this meeting, as well as the second and third NRC meetings and the field trial scorer training meetings were all conducted online rather than in-person.

Table 2.3: Test development processes and timeline

Year	Month	Group	Activity
2020	January	ICILS international study center	- Establishment of CIL and CT test specifications - Development of preliminary CIL and CT module concepts
2020	March	First meeting of national research coordinators (online)	- Reflections on ICILS 2018 tests of CIL and CT - Discussion of capturing process data in ICILS 2023 - Review of proposed preliminary CIL and CT module concepts
2020	April	ICILS international study center	- Revision of proposed preliminary CIL and CT module concepts following first meeting of national research coordinators - Development of five new CIL module storyboards and prototypes - Development of five new CT module storyboards and prototypes - Establishment of the ICILS test delivery software
2020	November - December	Second meeting of national research coordinators (online)	Review of the 10 new CIL and CT module storyboards and prototypes
2020	December	National research coordinators	- A web-based prioritization process following the second NRC meeting - NRCs ranked five CIL and five CT modules - Aggregated results were used by the ISC to prioritize modules for further development

Continued on next page



Table 2.3: Test development processes and timeline (Continued)

Year	Month	Group	Activity
2021	January	ICILS international study center	Revision of priority CIL and CT module storyboards and prototypes
2021	April	National research coordinators	- Web-based review of revised CIL and CT module storyboards and prototypes - Planning of CT pilot
2021	May	ICILS international study center	- Preparation of CIL and CT module content for software implementation - Pilot CT instrument preparation - CT piloting - Post-pilot review of CT instrument
2021	June - July	Third meeting of national research coordinators (online)	- Reflection on CT pilot - Review of modules proposed for inclusion in field trial test - Confirmation of test design
2021	July	ICILS international study center	- Revision of CIL and CT module content following third meeting of national research coordinators - Revision of test modules and development of content in online test delivery system
2021	July	ICILS international study center	Finalization of field trial test modules
2021	August	ICILS international study center and RM Results	Migration of test content to online delivery system
2022	March	ICILS field trial scorer training (online)	Review of field trial scoring guides for constructed-response items and large tasks (as part of scorer training)
2022	March - April	ICILS international study center	Revision of field trial scoring guides for constructed-response items and large tasks following scorer training

Continued on next page



Table 2.3: Test development processes and timeline (Continued)

Year	Month	Group	Activity
2022	April	ICILS international study center	• Analysis of field trial item data • Recommendations for modules/tasks to be included in main survey test (field trial analysis report)
2022	September	Fourth meeting of national research coordinators (Hamburg)	• Review of field trial analysis report • Recommendations for test design and modules/items proposed for inclusion in main survey
2023	March	ICILS main survey scorer training (Hamburg)	• Review of main survey scoring guides for constructed-response items and large tasks (as part of scorer training)
2023	March	ICILS international study center	• Finalization of main survey scoring guides for constructed-response items and large tasks

In ICILS, each task consists of several components: the stimulus materials presented to students, the questions or instructions given, the programmed behavior of the computer delivery system in response to students' actions, and the scoring logic. Tasks can be scored by humans, according to the scoring guides, or automatically, based on machine scoring specifications. The test development and review process covered all these elements within the ICILS modules.

The following lists present the main review questions used to guide reviews of CIL and CT tasks at all stages of the development process, from initial task development through to selection and finalization of field trial and main survey task content.

- Content validity
 - How does the material relate to the ICILS test specifications?
 - Do the tasks test the construct (CIL or CT) described in the assessment framework?
 - Do the tasks relate to content at the core of the aspects of the assessment framework or focus on trivial side issues?
 - How would the test content stand up to broader expert and public scrutiny?
- Clarity and fairness
 - Are the tasks and stimulus material coherent, unambiguous, and clear?
 - Are the modules and tasks interesting, worthwhile, and relevant?
 - Do the tasks assume prior knowledge and, if so, was this assumed to be acceptable or part of what the test intended to measure?
 - Is the reading load as low as possible without compromising the real-world relevance and validity of the tasks?
 - Are there idioms or syntactic structures that may prove difficult to translate into other languages?



- Do the constructed-response items and the large-task information provide clear guidance to students about what was expected in response to the items and tasks?
- Test-takers
 - Does the content of the modules and tasks match the expected range of ability levels, age, and maturity of the ICILS target population?
 - Does the material appear to be cross-culturally relevant and sensitive?
 - Are specific items or tasks likely to be easier or harder for certain subgroups in the target population for reasons other than differences in the ability measured by the test?
- Format and scoring
 - Is the proposed format the most suitable for the framework content being assessed by each task?
 - Is the key⁵ clearly correct and representing the ‘best’ answer to the question as asked?
 - Are the distractors⁶ plausible but also irrefutably and unequivocally incorrect?
 - Do the scoring criteria for the large tasks assess the essential characteristics of task completion?
 - Are there different approaches to provide answers with the same score, and did they represent equivalent or different levels of proficiency?
 - Is the proposed scoring consistent with the underlying ability measured by the test, and would test respondents with higher ability levels always score better than those with lower ones?
 - Are there other kinds of answers that have not been anticipated in the scoring guides (e.g., any that do not fall within the “correct” answer category description, but appear to be equally correct)?
 - Are the scoring criteria sufficient for scorers and do they clearly distinguish the different levels of performance?

Drafting and review of preliminary module ideas

Before the first meeting of the ICILS 2023 NRCs, the ISC developed preliminary concepts for the CIL and CT modules to be presented at the meeting. This meeting involved reflecting on the CIL and CT tests from ICILS 2018, discussing preliminary module concepts for inclusion in ICILS 2023, and considering the incorporation of process data into ICILS 2023. NRCs were invited to discuss and propose new ideas for module themes. The feedback on the preliminary content and suggested themes served as a foundation for subsequent module development.

Development of draft module storyboards

After the first NRC meeting, the ISC developed storyboards for 10 new modules (five for each of CIL and CT). The storyboards were created as interactive Adobe XD prototypes of each task in sequence. The tasks were presented in sequence so that those viewing the presentation could see the narrative sequence of tasks in the module.

Each Adobe XD layout contained the stimulus material with key elements designed to be interactive, together with the task instructions or question that students would be required to respond to. Each storyboard was accompanied by a set of implementation notes for each task or question. The notes elaborated the interactive elements and described the planned functionality/behavior of each task and provided instructions on how the task was to be scored. Instructions were provided not only for the human-scored tasks but also for the tasks that would be scored automatically by the computer system.

⁵ The correct answer to a multiple-choice question.

⁶ The incorrect options to a multiple-choice question.

**Presentation of draft module storyboards**

The second meeting of NRCs took place in November 2020 and was conducted online. As part of this meeting, the ISC presented the storyboards and interactive prototypes for consideration. When presenting the storyboards, NRCs made recommendations relating to the modules. The NRCs' feedback informed further revision of the module storyboards.

Web-based prioritization process

Following the second NRC meeting, a web-based prioritization process was conducted to rank the proposed CIL and CT modules. NRCs were provided with two forms, each listing five CIL and five CT modules, and were asked to rank them in order of preference. The aggregated rankings from all participating NRCs were used by the ISC to prioritize the most highly ranked modules for further development. As a result, seven module storyboards, three CIL and four CT, were selected for further development and refinement. Feedback from the second NRC meeting was incorporated into the refinement of these storyboards.

Interim review of prioritized module storyboards

After refining the prioritized module storyboards based on feedback from the second NRC meeting, the ISC held an online check-in meeting in April 2021. During this meeting, revisions to the tasks for the three CIL and four CT modules were presented.

Final review of proposed module storyboards

Following the online check-in meeting, the revised draft storyboards were presented at the third NRC meeting for final review. Feedback from this meeting was used to finalize the selection of modules for implementation in the online test delivery system for a pilot study and field trial, resulting in the decision to proceed with the three new CIL modules and four new CT modules.

Implementation of the final storyboards in field trial test delivery system and CT pilot study

The finalized storyboards were provided to RM Results for implementation into the online test delivery system. This process served two main purposes: (1) it allowed for the review and refinement of the draft modules based on their interactive features, and (2) it enabled the testing and enhancement of the test delivery system. A small number of countries participated in a CT pilot study from 24 May to 6 June, 2021. The pilot helped identify issues with the test delivery system and gathered student feedback on the difficulty of the CT tasks. This feedback, along with observations from the ISC, was presented to NRCs during the third NRC meeting.

Field-trial scorer training

National center representatives attended an online international scorer training meeting⁷ held before the field trial. These representatives subsequently trained the national center staff in charge of scoring student responses in their respective countries. Feedback from the scorer training process led to refinements to the scoring guides.

Field trial analysis review and selection of items for the main survey

Field trial data were analyzed to assess the measurement properties of the ICILS test items at both the international and national levels. Based on this analysis, the ISC recommended which modules and tasks should be included in the main survey instrument. These recommendations were discussed with NRCs at the fourth ICILS NRC meeting, where minor refinements were suggested for several tasks. The NRCs strongly recommended retaining all three new CIL test modules, and two of the four CT modules from the field trial for the main survey.

⁷ Separate online sessions were offered for the Northern and Southern Hemisphere countries.

**Post-field trial revision**

Minor modifications were made to a small number of tasks and the functionality of all tasks was further reviewed and refined. The main survey instruments, comprising seven CIL test modules (four trend modules, three newly developed modules) and four CT test modules (two trend modules, two newly developed modules), were then finalized.

Implementation of the final storyboards in the main survey online delivery system

Following the field trial, the module content was provided to RM Results to be authored into the main survey test delivery system. Further review and refinement of the content and function of the modules was conducted in this system.

Main survey scorer training

The main survey international scorer training meeting provided a final opportunity to reflect on the experience of scoring the field trial responses and to further review the scoring guides. The meeting was again attended by national center representatives who were responsible for training the national center staff in charge of scoring student responses in their respective countries. Feedback from the second international scorer training meeting along with student achievement data and the reported experiences of scorers during the field trial prompted further refinements to the scoring guides.

2.5 Field trial test design and content

The ICILS 2023 field trial tests of CIL and CT included both newly developed material and trend material which served two critical methodological purposes: equating new items with trend items to maintain scale continuity and joint analysis of item fit for new and trend items.

One of the primary objectives of ICILS is to report trends over time in CIL and CT. To achieve this, new test items introduced in ICILS 2023 must be linked to the existing CIL and CT scales used in previous study cycles. By including trend items (items that appeared in previous cycles) alongside new items, it is possible to estimate the difficulty of the new items relative to the established scale. This process, known as item equating, ensures that the measurement of student proficiency remains comparable across ICILS cycles.

Including trend material also allows for a more reliable psychometric evaluation of new items. In field trials, item fit analysis assesses whether test items function as expected within the measurement model. If only new items were analyzed, the results might be skewed due to uncertainty about the scale properties of entirely new content. However, by evaluating new and trend items together, we can assess how well the new items align with the CIL and CT constructs. Poorly fitting new items can then be identified and revised or excluded from the main study, ensuring the validity and reliability of the assessment.

Thus, trend items serve as anchors that enable both scale equating and robust psychometric analysis, ensuring that ICILS 2023 maintains consistency with previous cycles while introducing new assessment content effectively.

Test design

The CIL field trial test instrument included six test modules,⁸ yielding 160 score points from 112 scoring variables (see [Table 2.7](#)).⁹ Thirty-two percent of the CIL scoring variables were associated with information-based response tasks, 17 percent with skills tasks, and 51 percent with authoring tasks (see [Table 2.4](#)).

⁸ One trend module (School trip) was excluded from the field trial.

⁹ All authoring tasks and a small number of skills tasks were assessed using more than one criterion, with each criterion constituting a scoring variable.



Table 2.4: Field trial CIL test module composition: Number of scoring variables by task type

Module	Scoring variables by task type					Total
	Information-based response tasks		Skills tasks		Authoring tasks	
	Multiple choice	Constructed response	Linear skills	Nonlinear skills	Analytic criteria	
Board games club*‡	2	4	3	0	5	14
Breathing‡	0	2	3	0	9**	14
Computer use and health†	0	6	2	3	7	18
Internet safety†	3	4	1	0	15	23
Paper books vs ebooks†	2	5	1	1	9	18
Recycling***†	3	5	1	4	4 ^a , 8 ^b	25
Total	10	26	11	8	57	112

Note: * One linear skills item was excluded from the 2018 main survey and included in the 2023 field trial. One item classified as a nonlinear skills task in the 2018 main survey has been reclassified as a linear skills task for 2023.

** The contrast criterion was not scored in the 2023 field trial.

*** Three nonlinear skills items derived from a single task in the 2018 main survey were reported as a single item in Table 2.8 of the ICILS 2018 technical report. In this table they are reported as three nonlinear skills items.

a: Recycling simple authoring task

b: Recycling complex authoring task

† New module (developed in the current study cycle)

‡ Trend module (developed in a previous study cycle)

The CT field trial test instrument included six test modules and yielded 99 score points from 48 scoring variables (see Table 2.8). In total, 23 percent of CT scoring variables were associated with information-based response tasks, 27 percent with nonlinear systems transfer tasks, four percent with simulation tasks, and 46 percent to block-based coding tasks (see Table 2.5).



Table 2.5: Field trial CT test module composition: Number of scoring variables by task type

Module	Scoring variables by task type						Total
	Information-based response tasks		Nonlinear systems transfer	Simulation	Block-based coding		
	MC	CR			AC	AD	
Activity tracker†	0	0	2	1	4	1	8
Automated bus*‡	1	2*	4	1	0	0	8
Farm drone‡	0	0	1	0	5	3	9
Fingerprint matching†	0	1	2	0	1	2	6
Meal delivery†	1	4	1	0	2	0	8
Suns and moons†	0	2	3	0	0	4	9
Total	2	9	13	2	12	10	48

Note: * In Table 2.9 of the ICILS 2018 technical report the number of constructed response items in Automated bus was incorrectly reported as one and the two variables contributing to the single simulation item were counted separately.

MC: Multiple choice, CR: Constructed response, AC: Algorithm construction, AD: Algorithm debugging

[†] New module (developed in the current study cycle)

[‡] Trend module (developed in a previous study cycle)

A balanced complete block design was used for each of the CIL and CT tests. Each student completed a randomly assigned pair of CIL modules from the available six. In countries participating in the CT international option, students also completed a randomly assigned pair of CT modules from the six CT options. This design resulted in 30 possible combinations of CIL module pairs and 30 CT module pairs (see Table 2.6).

In countries administering only the CIL assessment, students were assigned one of 30 possible CIL module combinations, followed by one of two versions of the student questionnaire (see Chapter 4 for a description of the questionnaire development process). In countries participating in the CT option, students first completed one of the CIL permutations, then one of two questionnaire versions, and finally one of the 30 CT permutations.

A total of 120 different test forms were created, consisting of 30 CIL permutations, 30 CT permutations, and two versions of the student questionnaire. Each CIL and CT module appeared an equal number of times in the first and second module positions across the test forms, and each module was paired equally with the others. However, while all CIL were fully administered across all field trial countries, they were not fully balanced with the CT modules in terms of pairing frequency. This means that each CIL module did not appear an equal number of times with each CT module. This partial balancing was deemed sufficient because the tests of CIL and CT assessed different constructs, with both completed by only a subset countries, and students who completed both tests completed the questionnaire and had a break between the test of CIL and CT. The approach allowed for a manageable number of test forms in cases where both CIL and CT were administered.

Table 2.6: Field trial test form design and contents

CIL			CT		
CIL permutation	CIL module position		CT permutation	CT module position	
	1	2		1	2
1	B	G	1	FD	AB
2	B	R	2	FD	FP
3	B	I	3	FD	SM
4	B	C	4	FD	AT
5	B	P	5	FD	MD
6	G	B	6	AB	FD
7	G	R	7	AB	FP
8	G	I	8	AB	SM
9	G	C	9	AB	AT
10	G	P	10	AB	MD
11	R	B	11	FP	FD
12	R	G	12	FP	AB
13	R	I	13	FP	SM
14	R	C	14	FP	AT
15	R	P	15	FP	MD
16	I	B	16	SM	FD
17	I	G	17	SM	AB
18	I	R	18	SM	FP
19	I	C	19	SM	AT
20	I	P	20	SM	MD
21	C	B	21	AT	FD
22	C	G	22	AT	AB
23	C	R	23	AT	FP
24	C	I	24	AT	SM
25	C	P	25	AT	MD
26	P	B	26	MD	FD
27	P	G	27	MD	AB
28	P	R	28	MD	FP
29	P	I	29	MD	SM
30	P	C	30	MD	AT



Module name key

CIL modules

B: Breathing
G: Board games
R: Recycling
I: Internet safety (New CIL)
C: Computer use and health (New CIL)
P: Paper books vs ebooks (New CIL)

CT modules

AB: Automated bus
FD: Farm drone
FP: Fingerprint recognition (New CT)
SM: Suns and moons (New CT)
AT: Activity tracker (New CT)
MD: Meal delivery (New CT)

Field trial coverage of the CIL and CT constructs

All field trial tasks were designed and mapped according to the CIL and CT constructs. As stated in the ICILS assessment framework, "The ICILS tasks are designed to collect information about specific aspects of the relevant construct (CIL or CT), and each module typically includes content that addresses most, if not all, aspects of the construct. However, the test design of ICILS does not require equal representation of all aspects of the CIL and CT constructs" (Duckworth & Fraillon, 2024b, p. 76).

The CIL field trial instruments represented each strand of the CIL construct with varying levels of emphasis: 14 percent of the scoring variables corresponded to CIL Strand 1, 27 percent to CIL Strand 2, 39 percent to CIL Strand 3, and 20 percent to CIL Strand 4 (see Table 2.7). These proportions reflected the estimated time students were expected to spend on each CIL strand. CIL Aspects 2.1, 3.1, and 3.2 were assessed mainly through large tasks located at the end of each CIL module, with students expected to spend about two-thirds of their working time on these tasks. Overall, the test design adequately represented the four CIL strands.

Table 2.7: Field trial CIL scoring variable mapping to the CIL framework

	CIL framework aspect	Scoring variables (n)	Scoring variables (%)	Score points (n)	Score points (%)
1.1	Foundations of computer use	2	2	3	2
1.2	Computer use conventions	14	12	21	13
2.1	Accessing and evaluating information	22	20	33	21
2.2	Managing information	8	7	10	6
3.1	Transforming information	13	12	17	11
3.2	Creating information	30	27	44	28
4.1	Sharing information	14	12	19	12
4.2	Using information responsibly and safely	9	8	13	8
Total		112	100	160	100

In contrast, 71 percent of the scoring variables in the CT field trial instrument related to CT Strand 2, with the remainder covering CT Strand 1 (see Table 2.8). This is due to the nature of algorithm construction and debugging tasks, which, especially in the Farm drone module, were assessed using partial-credit scoring based on the accuracy and elegance of students' coding solutions. A detailed description of how the Farm drone module assessment tasks were scored is provided in Appendix D.

Table 2.8: Field trial CT scoring variable mapping to the CT framework

CT framework aspect		Scoring variables (n)	Scoring variables (%)	Score points (n)	Score points (%)
1.1	Knowing about and understanding digital systems	5	10	10	10
1.2	Formulating and analyzing problems	4	8	6	6
1.3	Collecting and representing relevant data	5	10	8	8
2.1	Planning and evaluating solutions	15	31	29	29
2.2	Developing algorithms, programs and interfaces	19	40	46	46
Total		48	100	99	100

Selection of CIL and CT test content for main survey

As stated previously, the field trial data were used to investigate the measurement properties of the ICILS test items, and ISC researchers made recommendations on which modules and tasks should be included in the main survey instrument. [Chapter 11](#) of this report describes the analysis procedures used to review the measurement properties. Following the recommendation from NRCs, the ISC decided to retain the three new CIL test modules and two of the four CT test modules for use in the main survey.

Also as mentioned previously, ISC researchers recommended refinements to a small number of tasks and to the scoring of some tasks. A task was only refined when data provided clear evidence of a problem with the task and if there was strong agreement that the refinement would very likely improve the task's measurement properties.

2.6 Main survey test design and content

Test design

The CIL main survey test instrument consisted of seven test modules (four trend modules and three new modules) with a total of 55 (see [Table 2.12](#)) tasks, yielding data from 108 scoring variables (see [Table 2.9](#)). Overall, 25 percent of the scoring variables were associated with information-based response tasks, 20 percent with skills task, and 55 percent with authoring tasks.

Some of these tasks generated several score points based on the criteria that were applied to the large tasks. The scoring variables shown in [Table 2.9](#) are those that were used in the scaling and analysis of the ICILS 2023 main survey CIL test data.

The CT main survey test instrument consisted of four test modules (two trend modules and two new modules) with a total of 31 tasks (see [Table 2.13](#)), yielding data from 31 scoring variables (see [Table 2.10](#)). The CT test emphasized block-based coding with 15 of the 31 scoring variables being associated with block-based coding tasks (see [Table 2.10](#)).

As was used in the field trial, the test modules were allocated using a complete balanced block design for each of CIL and CT. [Table 2.11](#) shows this design for the main survey. As described for the field trial, the term *test form* refers to each combination of modules used in the main survey.



Table 2.9: Main survey CIL test module composition: Number of scoring variables by task type

Module	Scoring variables by task type					Total
	Information-based response tasks		Skills tasks		Authoring tasks	
	Multiple choice	Constructed response	Linear skills	Nonlinear skills	Analytic criteria	
Board games club [‡]	2	4	2	0	5	13
Breathing [‡]	0	3	3	0	10	16
Computer use and health [†]	0	2	1	3	8	14
Internet safety [†]	2	3	1	0	10	16
Paper books vs ebooks [†]	1	2	1	1	8	13
Recycling [‡]	2	5	2	2	4 ^a , 8 ^b	23
School trip [‡]	1	0	4	2	7	14
Total	8	19	14	8	60	109

Note:

a: Recycling simple authoring task

b: Recycling complex authoring task

[†] New module (developed in the current study cycle)[‡] Trend module (developed in a previous study cycle)

Table 2.10: Main survey CT test module composition: Number of scoring variables by task type

Module	Scoring variables by task type					Total
	Information-based response tasks		Nonlinear systems transfer	Simulation	Block-based coding	
	MC	CR			AC AD	
Activity tracker [†]	0	0	2	1	3 1	7
Automated bus* [‡]	1	2*	4	1	0 0	8
Farm drone [‡]	0	0	1	0	5 3	9
Suns and moons [†]	0	1	3	0	0 3	7
Total	1	3	10	2	8 7	31

Note: * In Table 2.9 of the ICILS 2018 technical report the number of constructed response items in Automated bus was incorrectly reported as one and the two variables contributing to the single simulation item were counted separately.

MC: Multiple choice, CR: Constructed response, AC: Algorithm construction, AD: Algorithm debugging

[†] New module (developed in the current study cycle)[‡] Trend module (developed in a previous study cycle)



Table 2.11: Main survey test form design and contents

CIL			CT		
CIL permutation	CIL module position		CT permutation	CT module position	
	1	2		1	2
1	G	R	1	FD	AB
2	G	I	2	FD	SM
3	G	C	3	FD	AT
4	G	P	4	AB	FD
5	G	B	5	AB	SM
6	G	S	6	AB	AT
7	R	G	7	SM	FD
8	R	I	8	SM	AB
9	R	C	9	SM	AT
10	R	P	10	AT	FD
11	R	B	11	AT	AB
12	R	S	12	AT	SM
13	I	G			
14	I	R			
15	I	C			
16	I	P			
17	I	B			
18	I	S			
19	C	G			
20	C	R			
21	C	I			
22	C	P			
23	C	B			
24	C	S			
25	P	G			
26	P	R			
27	P	I			
28	P	C			
29	P	B			
30	P	S			
31	B	G			
32	B	R			
33	B	I			
34	B	C			

Continued on next page



Table 2.11: Main survey test form design and contents (Continued)

CIL			CT	
CIL permutation	CIL module position		CT permutation	CT module position
35	B	P		
36	B	S		
37	S	G		
38	S	R		
39	S	I		
40	S	C		
41	S	P		
42	S	B		

Module name key**CIL modules**

G: Board games
 R: Recycling
 I: Internet safety (New CIL)
 C: Computer use and health (New CIL)
 P: Paper books vs ebooks (New CIL)
 B: Breathing
 S: School trip

CT modules

FD: Farm drone
 AB: Automated bus
 SM: Suns and moons (New CT)
 AT: Activity tracker (New CT)

Main survey coverage of the CIL and CT constructs

All main survey CIL tasks were developed according to and mapped against the CIL and CT constructs. Table 2.12 shows this mapping.

Table 2.12: Main survey CIL scoring variable mapping to the CIL framework

	CIL framework aspect	Scoring variables (n)	Scoring variables (%)	Score points (n)	Score points (%)
1.1	Foundations of computer use	3	3	3	2
1.2	Computer use conventions	11	10	15	10
2.1	Accessing and evaluating information	18	17	22	14
2.2	Managing information	9	8	11	7
3.1	Transforming information	15	14	22	14
3.2	Creating information	32	29	54	35
4.1	Sharing information	11	10	12	8
4.2	Using information responsibly and safely	10	9	17	11
Total		109	100	156	100

The 109 main survey scoring variables yielded 156 score points for inclusion in the item analysis and scaling. Overall 12 percent of score points were derived from scoring variables corresponding to CIL Strand 1, 21 percent to CIL Strand 2, 49 percent to CIL Strand 3, and 19 percent to CIL Strand 4.

A comparison of Table 2.7 and Table 2.12 reveals that the final test instrument provided very similar CIL framework coverage to that of the field trial instrument.

All main survey CT tasks were developed according to and mapped against the ICILS CT framework. Table 2.13 shows this mapping.

Table 2.13: Main survey CT scoring variable mapping to the CT framework

	CT framework aspect	Scoring variables (n)	Scoring variables (%)	Score points (n)	Score points (%)
1.1	Knowing about and understanding digital systems	4	13	9	14
1.2	Formulating and analyzing problems	2	6	4	6
1.3	Collecting and representing relevant data	4	13	7	11
2.1	Planning and evaluating solutions	9	29	19	29
2.2	Developing algorithms, programs and interfaces	12	39	26	40
Total		31	100	65	100

The 31 main survey scoring variables yielded 65 score points for inclusion in the item analysis and scaling. Overall 31 percent and 69 percent of score points were derived from items corresponding to CT Strands 1 and 2 respectively.



A comparison of Table 2.8 and Table 2.13 reveals that the final test instrument provided similar CT framework coverage to that of the field trial instrument. The relative decrease in CT Strand 2.1 and 2.2 coverage was the result of removing tasks that were identified as being too difficult for the target population.

2.7 Released test modules

Two CIL (Breathing, School trip) and two CT (Automated bus, Farm drone) modules (see Table 2.1 and Table 2.2), have been released since publication of the ICILS international report (Fraillon, 2025).

Online demonstration versions of each of these four modules can be accessed at: <https://www.iea.nl/icils2023-assessment>

Detailed descriptions of these tasks are provided in Appendix E and Appendix F of the ICILS 2023 international report (Duckworth, 2025a, 2025b).

Appendix C and Appendix D provide summary information about the human and machine scoring rules used to score students responses for all released CIL and CT tasks. In addition to the scoring rules, an explanation of the scoring methods used for the Farm drone module are provided in Appendix D.

See Appendix G for information about how the scoring variables were recoded during scaling and analysis.

References

- Duckworth, D. (2025a). Appendix E: CIL release modules. In J. Fraillon (Ed.), *An international perspective on digital literacy: Results from ICILS 2023* (pp. 255–269). Springer. <https://doi.org/10.1007/978-3-031-87722-3>
- Duckworth, D. (2025b). Appendix F: CT release modules. In J. Fraillon (Ed.), *An international perspective on digital literacy: Results from ICILS 2023* (pp. 271–285). Springer. <https://doi.org/10.1007/978-3-031-87722-3>
- Duckworth, D., & Fraillon, J. (2024a). Computational thinking framework. In J. Fraillon & M. Rožman (Eds.), *IEA International Computer and Information Literacy Study 2023 assessment framework* (pp. 35–43). Springer. https://doi.org/10.1007/978-3-031-61194-0_3
- Duckworth, D., & Fraillon, J. (2024b). ICILS instruments. In J. Fraillon & M. Rožman (Eds.), *IEA International Computer and Information Literacy Study 2023 assessment framework* (pp. 57–81). Springer. https://doi.org/10.1007/978-3-031-61194-0_5
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (Eds.). (2020). *IEA International Computer and Information Literacy Study 2018 technical report*. International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>
- Fraillon, J., Schulz, W., Friedman, T., Ainley, J., & Gebhardt, E. (Eds.). (2015). *IEA International Computer and Information Literacy Study 2013 technical report*. International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2013-technical-report>
- Fraillon, J. (Ed.). (2025). *An international perspective on digital literacy: Results from ICILS 2023*. Springer. <https://doi.org/10.1007/978-3-031-87722-3>
- Fraillon, J., & Duckworth, D. (2024). Computer and information literacy framework. In J. Fraillon & M. Rožman (Eds.), *IEA International Computer and Information Literacy Study 2023 assessment framework* (pp. 21–34). Springer. https://doi.org/10.1007/978-3-031-61194-0_2
- Fraillon, J., & Rožman, M. (2024). *IEA International Computer and Information Literacy Study 2023 assessment framework*. Springer. <https://doi.org/10.1007/978-3-031-61194-0>

Chapter 3:

ICILS survey operations software

Sebastian Meyer and Daniel Duckworth

3.1 Introduction

This chapter outlines the key features of the survey operations software employed in ICILS 2023 and describes selected challenges associated with using the software in a cross-national, large-scale assessment. The chapter begins by describing the individual survey operations software applications and systems, as well as their integration with other technical systems used in survey operations. Further details of specific procedures and related topics are discussed in the following chapters: assessment and test design in [Chapter 2](#), translation and adaptation in [Chapter 5](#), field operations in [Chapter 8](#), data flow and integration in [Chapter 10](#), and scaling and analysis of the test materials in [Chapter 11](#).

3.2 Survey operations software applications and systems

IEA authored applications

IEA-developed survey operations software applications and systems used in ICILS 2023 included the following:

- Windows Within-School Sampling Software (WinW3S): This application helps countries manage within-school sampling and test administration procedures.
- Online Survey System (OSS): This system supports the translation, adaptation, and delivery of computer-based questionnaire materials.

Both WinW3S and OSS were customized for use in ICILS 2023 survey operations.

Use of WinW3S was mandatory in all participating countries, primarily because the application ensures adherence to the within-school teacher and student sampling procedures stipulated by the study's design. All countries were required to import the school sample drawn by IEA into WinW3S and list all eligible teachers and target grade classes. The software then randomly selected the teacher and class samples within each school. Any modifications to the sample selected by the application were strictly prohibited. However, exceptions were allowed in cases of listing errors that were identified after school sample selection, but these had to be discussed with the IEA sampling team. All changes were documented in the underlying database. After survey administration, countries were required to enter participation data for all sampled teachers and students within selected classes. The participating countries were required to provide the database containing all sampling and tracking information to IEA after data collection. For more information on the within-school sampling design, see [Chapter 6](#).

Data at the school (principal and ICT coordinator) and teacher levels were collected using the OSS across all participating countries. Designed as a comprehensive solution for survey instrument authoring, data collection, and monitoring, the OSS is structured into three distinct components to accommodate various usage scenarios. The Designer component facilitated the creation, deletion, disabling, and editing of survey elements (e.g., questions and response categories) and their properties. It allowed for the translation of all text passages from the questionnaires, as well as additional system texts, ensuring consistency across languages. The Web component provided respondents with access to questionnaires in an HTML format, allowing completion directly within standard internet browsers. Lastly, the Monitor component, a web-based tool, enabled national centers to track participation in real time, facilitating follow-ups with schools when questionnaires were incomplete or missing.



Commercially provided applications

In addition to the survey operations software provided by IEA, ICILS 2023 also employed the RM Assessment Master (AM) software system. The AM software system facilitated the production and delivery of the computer-based assessment content, as well as the collection of student data and the scoring of student test response data. The AM modules used in ICILS 2023 included the following:

- AM Manager: An administration module for managing countries, schools, students, and users;
- AM Designer: An item authoring, test form design and translation module;
- AM Examiner: The instrument delivery engine; and
- AM Marker: A scoring module.

3.3 Procedures and software user manuals

A comprehensive series of manuals, known as Survey Operations Procedures (SOP) manuals, guides the work of national center staff during IEA studies. These SOP manuals, developed through close collaboration between IEA and the International Study Center (ISC) team, contain detailed instructions and guidelines for the use of essential software systems, including WinW3S and OSS. Each manual is specifically tailored to meet the needs of the project, ensuring that national center staff can efficiently manage operations and maintain consistency across all participating countries (see [Chapter 8](#) for details).

For ICILS 2023, the following software user manuals were incorporated into the SOP materials:

- *IEA WinW3S Manual*: This manual provided detailed guidance on using the IEA Windows Within-School Sampling Software. It outlined procedures for tracking student and teacher participation, completing sampling forms, and coordinating test sessions at the school level.
- *IEA OSS Manual*: The IEA Online Survey System manual supported national centers in adapting questionnaire materials, entering translations, and managing the online survey administration. Detailed steps for checking translations, adaptation tracking, and finalizing questionnaires for field operations were provided.
- *Delivery Engine Module (AM Examiner)*: For ICILS 2023, the manuals related to the delivery engine module (AM Examiner) were included within the instructions provided to school coordinators and test administrators. These supplementary manuals provided step-by-step guidance on preparing for the ICILS compatibility checks to be run in schools so that the computers intended to be used for the IEA ICILS 2023 main survey were suitable and set up properly to run the student assessment, setting up computers for running the tests, administering the tests, and troubleshooting issues during test sessions. They also included protocols for managing the different test delivery methods (using the internet, USB, or a local server).
- *RM AM Manual*: The RM AM manual supported national centers in translating and reporting layout issues for the test instruments, implementing feedback from translation and layout verification, and preparing and administering the study instruments. A dedicated section in the manual was provided for the translation module within the AM system. This section guided national center staff through the translation and adaptation of student test items, ensuring that the translated content was accurate and aligned with the international source version. The AM Manager module enabled the management of student test sessions.
- *Scoring Manual (AM Marker)*: This module enabled the assignment of scoring tasks and scoring student responses.

These manuals were reviewed after the field trial data collection and updated to incorporate feedback from national centers.

3.4 ICILS 2023 system architecture

ICILS 2023 employed a combination of survey operations software applications and systems to support the study's preparation, administration, and post-processing phases. [Table 3.1](#) provides an overview of the applications and systems with their associated instrument development, delivery, and data collection processes. The table shows how these components interfaced and interacted.

Table 3.1: ICILS survey operations software applications and systems

Process	Instruments	
	Student test and questionnaire	Principal, ICT coordinator, and teacher questionnaires
Authoring international source instruments	AM Designer (authoring from storyboards and functional specification)	IEA OSS admin module (authoring from document files)
Translation	AM Designer	IEA Translation System
Sample selection and ID provisioning	IEA WinW3S	
Delivery systems	AM Examiner	IEA OSS delivery module
Initialization	Student ID, password, and language information from IEA WinW3S	Respondent ID and password from IEA WinW3S
Recommended delivery method(s)	Internet browser-based; USB drives	Internet browser-based: self-administered
Alternative delivery method(s)	Local server or carry-in laptop sets (where school infrastructure was insufficient)	N/A
Data capture	Directly into central database; directly onto USB drives; alternatively, local laptop servers (student by student)	Directly into central database (all respondents)
Data merging (across respondents)	AM Manager (uploaded to)	N/A
Scoring	AM Marker; AM Examiner (machine scoring)	N/A
Data management	IEA WinW3S (crosscheck of expected versus exported data)	

3.5 Instrument preparation

The student instruments (test and questionnaire) were authored, adapted and translated in the AM Designer module. The teacher, ICT coordinator and principal questionnaires were authored, and adapted and translated text was entered using the IEA OSS. In this section, the procedures used for each software system are described separately. The key for using two different systems is that questionnaires focus on survey-based self-reported data collection, while student assessments require an environment capable of delivering structured performance-based tasks with automated scoring. Using OSS for questionnaires and Assessment Master for student instruments allowed ICILS to leverage the most suitable platform for each type of data collection, ensuring both data quality and international comparability.

Authoring

Student instruments

The ICILS 2023 student questionnaire was authored directly in the RM AM system using the authoring tools within the AM Designer module. No new development was required to support the questionnaire development and delivery.

The development of the features within the test delivery platform specific to the ICILS 2023 student test occurred in parallel with the test development process (see [Chapter 2](#) for details on test development). The six trend test modules—two computer and information literacy (CIL) from ICILS 2013, two CIL and two computational thinking (CT) modules from ICILS 2018—were already available and largely ready to use. However, to accommodate changes in the delivery system, RM Results updated certain backend functionalities in the AM Examiner module. These updates ensured that the software continued to function effectively and did not alter the user experience of the trend modules compared to previous cycles, thus ensuring consistency for participating students across cycles.

For ICILS 2023, three new CIL modules and two new CT modules were developed. After the initial storyboards for these new modules were completed, they were sent to RM Results for implementation into the delivery platform. This process involved multiple iterations to review, refine, and test the functional aspects of each task to ensure that the intended tasks were fully operational.

Principal, ICT coordinator, and teacher questionnaires

The ICILS 2023 principal, ICT coordinator, and teacher questionnaires were authored directly in the IEA OSS system using the authoring tools within the OSS Designer module. The OSS Designer provided a flexible framework for constructing a variety of item types and questionnaire structures tailored to the different respondent groups.

The international versions of the questionnaires were developed by the ICILS ISC and implemented into the OSS platform without requiring additional software development. The system allowed for seamless integration of the finalized questionnaire content into a survey environment suitable for previewing and national adaptation.

Once the questionnaires were authored, IEA compiled an ICILS 2023-specific OSS software package containing the finalized international versions. This package was distributed to national centers, where it could be installed on local systems. The installation included a fully functional local web server, enabling national center staff to input verified translations and apply pre-approved national adaptations. This setup allowed users to preview the questionnaires in an environment closely resembling the actual delivery system, supporting thorough review and quality assurance prior to final deployment.

Adaptation

Two types of adaptation were possible for the ICILS instruments.

Structural adaptations were possible for the questionnaires but not for the student test. Structural adaptations include country-specific alterations to the structure of a question. These could be, for example, the addition of some national items within an international question, the addition of an entirely new question, or creation of additional response categories (that are sub-categories of the original international response category).

Language adaptations involved the replacement of specific words or phrases in an instrument with locally equivalent terms. This differs from translation in that, in some instances, a direct translation of a word or phrase in the source language may not convey the intended meaning in the target language. Terms or phrases requiring language adaptation were shown in the instruments in square brackets []. Once language adaptations requested by a country were approved by the ISC, they were included. The national center team entered the text into the translated version of their national instrument using the same method as entering all other translated text (see [Chapter 5](#) for more details on the adaptation procedures).



Student instruments

Structural adaptations to the national version of the student questionnaire had to be custom-coded into the instrument for each country by the ISC staff. Once a country's requested adaptations were approved by the ISC, the ISC team implemented them in AM Designer. For each adaptation, the original questionnaire item was copied into a national variable, where the agreed-upon structural changes were applied. This approach clearly flagged the item as a nationally adapted version. The original item was then hidden from the administered questionnaire. After implementation, national centers reviewed and verified the adaptations to ensure accuracy and alignment with their national requirements.

Principal, ICT coordinator, and teacher questionnaires

Structural adaptations to the national versions of the principal, ICT coordinator, and teacher questionnaires had to be implemented by national center staff in accordance with the guidelines outlined in the IEA OSS manual. The system strictly prohibited structural modifications to any existing international question or question part. To implement adaptations, users were required to create a copy of the relevant question, apply the necessary structural changes, and ensure that the original international version was hidden from the administered questionnaire. As with the student questionnaire, this approach clearly flagged the item as a nationally adapted version. The web-preview functionality allowed users to verify that the structural adaptations were correctly implemented before finalizing the questionnaire.

Translation

The translation processes used in AM Designer (for the student instruments) and OSS Designer (for the principal, ICT coordinator, and teacher questionnaires) differed substantially. The main distinction was that, for the student instruments, the entire translation and verification process was conducted within the AM Designer system. In contrast, for the questionnaires for principals, ICT coordinators, and teachers, translation and verification were completed outside the OSS system. As a result, the following descriptions of the two environments—AM Designer and OSS Designer—differ accordingly.

Student instruments

In ICILS 2023, the translation system within the AM Designer module was used to translate the student test and student questionnaire. Every text element in the student instruments required translation, with the system providing an empty text field for users to enter the corresponding translation. A preview functionality allowed users to verify whether all text elements had been translated correctly. In addition to this core functionality, the translation system offered the following features:

- Some selective functionality relating to the role of the user (e.g., administrator, translator, translation reviewer, translation verifier, verification reviewer);
- Opportunity to enter translated text for all screen elements;
- Ability for those countries where the test would be administered in more than one language to select the target language;
- Ability to view the translation history of a text element;
- Ability to view the source text, the translated text, and any previous revisions of the translated text, and to compare translated versions;
- A comments interface to allow translators, reviewers, and verifiers to enter comments linked to elements of text;
- Opportunity to view “live” translated versions of the tasks at any point during the translation and to simultaneously view (in a separate window) a live version of the source task;
- Ability to enter text as plain text, including editable HTML tags or to enter and use formatting tools to format text;



- Ability to view translated text elements as plain text or as rendered HTML;
- Ability to search for and/or selectively bulk-replace text within tasks;
- Ability to bulk-import the field trial translations so that they could be used as a starting point for main survey translations;
- Opportunity for translators, reviewers, and verifiers to monitor the progress of task completion (i.e., translation state); and
- Facility for designated users to add or remove questions and items for questionnaire instruments (structural adaptations).

For more information on the translation procedures please refer to [Chapter 5](#).

Principal, ICT-coordinator, and teacher questionnaires

The translation of the principal, ICT coordinator, and teacher questionnaires was carried out by national centers outside the OSS system. National centers received Microsoft Word files containing the international source versions of the questionnaires for translation and verification (see [Chapter 5](#) for more details on the translation procedures). Once the translated content had been finalized and verified, national centers manually entered the translations into the OSS Designer for each administered language.

The OSS Designer offered several features to support this process:

- A dedicated text field for each translatable element in the international source version;
- The ability to input translations manually for each item and screen;
- A built-in check functionality to verify that all required translations were entered;
- A web preview mode to confirm that translated text displayed correctly in the final questionnaire format, reflecting how it would appear to respondents; and
- Support for multiple administered languages.

For more information on the translation procedures please refer to [Chapter 5](#).

3.6 Student instrument delivery using AM Examiner

The student instrument delivery engine provided the following features:

- Delivery of multiple choice, constructed response, and drag and drop items;
- Delivery of linear and nonlinear skills tasks (based on software simulations with real-world responsiveness);
- Delivery of authoring tasks (live software applications with functionality to add, edit, and format multimodal media);
- Display of closed web environments with the facility to display multiple pages and navigate between and within pages with a simulated web browser;
- Delivery of nonlinear systems transfer and simulation tasks;
- Delivery of block-based coding tasks;
- Capture of all student final responses to each task;
- Capture of time taken on each task;
- Capture of event data on each task;



- Facility to codify and machine score student responses to specified tasks;
- A progress display (number of tasks completed and to come per module) for students; and
- A countdown timer for students per module.

Regarding the data capture, several mechanisms were implemented to ensure detailed, structured, and standardized output. Final responses were saved periodically and stored in the native RM system, linked to unique identifiers for each student, module, and task. These responses were later transformed into flat files—where each row represented a student and each column represented that student's response to a specific item—in CSV format for further processing. Time-on-task data were captured in seconds, measuring the duration from task load to either proceeding to the next task or exiting the module. For selected tasks, detailed event data were collected, including timestamped logs of user interactions such as clicking named user interface elements, typing, and navigation actions.

In order to maximize participation, the ISC worked with RM results to ensure that the delivery could be as flexible as possible, allowing for two main forms of delivery of the student instruments:

- Internet (browser-based) delivery; and
- USB flash drive delivery for use with school computers running Windows operating systems.

The test delivery engine itself was web-based. Offline solutions made use of a locally installed web server (on USB or on the local server) through which the application could be accessed. Although the USB-based delivery engine was Windows-based, it could be run using some Windows emulation software on Mac OS X or Linux.

All offline solutions used a portable version of the Mozilla Firefox browser. All ICILS 2023 materials were developed to be compatible with recent versions (not more than one year old) of the following browsers:

- Mozilla Firefox
- Google Chrome
- Microsoft Internet Explorer 11 (IE11)
- Microsoft Edge
- Apple Safari
- Opera

AM Designer, AM Marker, and AM Manager functioned on recent versions of all major browsers. As part of the field-operation procedures, data collected using offline solutions (USB or local server) needed to be uploaded to a central database as soon as practicable after each test session. Web-based delivery did not require any additional steps, as the data were immediately written to the AM database during each test session.

Two backup delivery alternatives were also available to countries:

- Taking laptop computers with AM Examiner already installed into schools; or
- Connecting a single laptop computer into the school network to act as a web server that hosts the student instruments (this was recommended for use with schools using networked computers that did not use Windows operating systems such as Mac OS or Linux). This method is hereafter referred to as the “local server method.”

**Recommended delivery methods**

Internet delivery and USB delivery (in schools where internet delivery was not feasible) were the two recommended delivery methods for use in ICILS 2023.

Internet (browser-based) method

The student instrument delivery engine, AM Examiner, was web-based, which meant that the instruments could technically be delivered over the internet and accessed using a web-browser. In ICILS 2018, operational constraints led to the decision to primarily administer the instruments using USB drives. These USB drives acted as local web servers, containing the test and questionnaire content, and were used to administer the instruments on individual computers. Although the delivery engine was designed to run on Windows, it could also operate on Mac OS X or Linux using certain forms of Windows emulation. In ICILS 2023, however, internet-based delivery was offered to all participating countries, expanding the flexibility of the instrument administration. Schools with adequate infrastructure could now deliver the instruments directly through an internet browser. Where this method was used, student data were uploaded in real-time to the central ICILS database.

USB flash drive method

The USB flash drive method remained available for schools with limited internet connectivity. The use of a portable version of the Mozilla Firefox browser, hosted on each USB drive, enabled a consistent experience across different systems. All ICILS 2023 materials were developed and tested using several browsers to ensure compatibility across browsers.

Where this method was used, after completion of each test session, test administrators were required to complete the additional steps of locating a computer with an internet connection and then uploading the student data from each individual USB drive to the central ICILS database.

Alternative delivery methods**Providing computers to schools**

In schools where the computers did not meet specifications, national centers were asked to consider using the local network server method, or to provide computers to the school (such as a set of laptop computers) that could then be configured for use with any of the three available delivery methods.

Local network server method

An alternative (but not recommended) delivery method was the use of a local network server. Under the local network server method, the delivery engine was installed on a notebook computer, which was then connected to the school's local area network (LAN). Through this setup, students could access the test content locally using a modern web browser (e.g., IE11, Edge, Opera, Firefox, Chrome, Safari), without requiring individual USB drives for each computer. This option was designed to support schools where internet connectivity was unstable or where USB ports were disabled for security purposes but a reliable internal network was available.

Where this method was used, after completion of each test session, test administrators were required to complete the additional steps of connecting the local server computer to the internet and then uploading the student data from the test session to the central ICILS database. However, the local network server method proved to be problematic for countries that chose this option, which is discussed in more detail in section 3.10.

Choosing a delivery method

In order to support national centers and school coordinators to choose the most suitable delivery method within each school, and to maximize the chances of successful and smooth test delivery, school coordinators were asked to complete a questionnaire describing the available instrument delivery infrastructure (e.g., hardware, internet provision, and software security controls) and were provided with specific minimum specifications for the computers used to administer the instruments for each method.



As part of this process, schools were also asked to run simple compatibility checks of the internet- and USB-based methods. The internet-based method check required the user to navigate and log in to a website. They received immediate feedback of whether or not the access was successful and whether or not the necessary configurations were in place to successfully run the testing software. The USB-method test was similar, except that the users were required to run the testing software from a USB drive. In this test, the user was further asked to execute a simulated data upload, which also determined whether data could be successfully written to the USB drive. NRCs, in collaboration with school coordinators, were responsible for using data from the questionnaire and the results of the compatibility tests to determine the best delivery method within schools. In many countries a single delivery method was used (e.g., internet- or USB-based) although many countries using the internet-based delivery method made provisions to use the USB-delivery method as a backup.

Following are the minimum specifications provided to schools regarding the successful implementation of each delivery method.

Internet (browser-based) method

The internet (browser-based) method required each student to have an internet connected and compatible computer that accessed the *ICILS Student Software* from a cloud server via a web browser. This method was more flexible than the USB method because it could be used on computers with any operating system (Windows, Mac, Linux) as long as they had a modern web browser (updated within the past calendar year) and a stable internet connection. Typically, school computers required little or no configuration under this method. In some cases security settings may have needed to be altered to allow access to the ICILS testing URL.

School coordinators were provided with the following list of minimum specifications for computers to run the ICILS student test using the internet (browser-based) method.

- Screen resolution: A minimum resolution of 1280x800 pixels was required to ensure that all test items displayed correctly.
- Display size: A minimum display size of 11-inches was required to ensure rendered content was large enough to see.
- Operating system: Any operating system that can run a modern web browser (e.g., IE11, Edge, Opera, Firefox, Chrome, Safari).
- Peripherals: A mouse and external keyboard was required to ensure students could easily engage with complex interactive assessment tasks.
- Power supply: Laptops were required to use a power supply to ensure an uninterrupted test experience.

USB flash drive method

Under the USB flash drive method, the ICILS student instrument software was stored on individual USB flash drives. The test was delivered to students by plugging one USB flash drive into each computer used by a student and running the software from the USB flash drive. The computers could be local school computers or external computers provided for the ICILS administration. This method was limited to computers running Windows 8 or higher.

Features of the USB flash drive method were:

- The whole *ICILS Student Software* system was contained on the USB flash drive.
- No local network or internet connection was required to complete test sessions.
- Results were uploaded from the USB flash drive when connected to the internet.



- A separate USB flash drive containing the *ICILS Student Software* was required for each individual student in each test session. One USB flash drive could be used by multiple students across different test sessions.

School coordinators were provided with the following list of minimum specifications for computers and USB drives to run the ICILS student test using the USB flash drive method.

- Screen resolution: A minimum resolution of 1280x800 pixels was required to ensure that all test items displayed correctly.
- Display size: A minimum display size of 11-inches was required to ensure rendered content was large enough to see.
- Operating system: Windows 8 (with the latest patches).
- Peripherals: A mouse and external keyboard was required to ensure students could easily engage with complex interactive assessment tasks.
- Power supply: Laptops were required to use a power supply to ensure an uninterrupted test experience.
- USB flash drive brand: Recommended brands were Samsung, SanDisk, Kingston, PNY, Corsair to ensure file integrity.
- USB capacity: A minimum storage capacity of 16GB was required to store the *ICILS Student Software*.

Local network server method

The local server method was a hybrid of the USB and internet delivery methods. This approach was used when internet access was unreliable but the school had a functioning internal network. The ICILS student instrument software was loaded on a single server computer and then individual student (client) computers connected to the computer through a local area network (LAN). The student experience of the test was equivalent to that during internet delivery, and the client computers had the same requirements as those using internet delivery. However, all student data in each test session were saved locally on the server computer and subsequently had to be uploaded immediately after the test session.

School coordinators were provided with the following list of minimum specifications for server computers to run the ICILS student test using the local server method.

- Operating system: Windows 8 (with the latest patches).
- Server CPU: 2.2 GHz processor.
- Memory capacity: 8 GB RAM.
- Storage capacity: 10 GB of available HDD storage.
- Privileges: Administrator level user privileges.

The minimum computer requirements for the computers allocated to students were the same as the internet (browser-based) method. If the server was using a WiFi hotspot, the allocated computers also needed a WiFi device.

3.7 Principal, ICT coordinator, and teacher questionnaires delivery using OSS web component

The OSS web delivery engine was used to administer the ICILS 2023 questionnaires for principals, ICT coordinators, and teachers. Unlike the student instruments, the OSS component supported only a



single delivery mode—online browser-based delivery—which allowed respondents to access and complete their questionnaires using an internet browser. Despite the streamlined setup, the OSS system offered a robust and flexible platform for questionnaire administration.

The OSS delivery engine supported the following question formats:

- **Question formats:**

- Single (non-categorical);
- Single (categorical);
- Single (check all that apply);
- List (non-categorical);
- Matrix (categorical);
- Matrix (check all that apply); and
- Mix Matrix (combination of categorical and non-categorical in a matrix).

In addition, the system provided the following features to support administration and data capture:

- **Features**

- Capture of all final responses to each question;
- Capture of process data on the questionnaire, including timestamps for initial and final logins; and
- A progress display (completion status of each question and question part).

Online-only delivery

Questionnaires were delivered exclusively online and accessed via a secure web browser interface. Respondents were provided with unique credentials to log in and complete their assigned questionnaire. The system supported a responsive layout, ensuring accessibility across a wide range of screen sizes and resolutions.

Given the relatively lightweight nature of the questionnaire interface, minimum hardware requirements were modest. Computers were required to have a Pentium processor (1 GHz or higher), 512 MB of RAM or more, and one of the following supported browsers:

- Google Chrome;
- Mozilla/Firefox 20.0 and above;
- Microsoft Edge;
- Microsoft Internet Explorer 7.0 and above;
- Apple Safari/WebKit 1.0 and above; or
- Opera 7.0 and above and Konqueror (all versions regardless of platform).

Operational procedures

Before the start of data collection, national centers created access credentials via WinW3S system. Each questionnaire was activated by the ISC within the system according to the national data collection and instrument production schedule. National centers were responsible for distributing access credentials and providing basic technical support to respondents where needed.

To ensure smooth administration, national centers were advised to verify that respondents had access to a computer that met the minimum requirements and could successfully access the questionnaire



interface. In contrast to the student instruments, no compatibility check tools or offline alternatives were required, significantly reducing the technical complexity of administration.

3.8 Student response scoring and coding

Scoring student test responses using AM Marker and AM Examiner

The ICILS student test included both human-scored and automatically scored tasks. Automatically scored tasks followed a set of machine scoring rules established by the ISC. RM integrated the scoring logic into the AM Examiner module, while the ISC defined and evaluated multiple test cases for each machine-scored task to ensure the implementation aligned with expectations.

AM Marker

The ICILS 2023 scoring module, AM Marker, was adapted from the ICILS 2018 scoring module. It provided the following key features:

- Role-based access for different user types (e.g., administrator, team leader, scorer).
- Blind double-scoring functionality to ensure marker reliability by requiring a subset of student responses to be scored independently by multiple scorers.
- Early scoring capability, allowing scoring to begin before all student responses were uploaded, without compromising the double-scoring procedure.

Scoring workflow and user roles

To facilitate scoring, three primary user roles were established:

- Scorer – Engaged in training with pre-scored student responses and scored student work.
- Scoring Team Leader – Had all scorer privileges plus full access to scoring and management functions.
- Admin Scorer – Included all team leader and scorer functions, with additional permissions to create scorer accounts and manage scoring, training, and administrative tasks in AM Marker.

Additionally, NRC users had all admin scorer permissions plus full access and control over the AM Marker module.

Role-specific permissions

Scorers could:

- View test tasks as presented to students, alongside student responses.
- Assign scores to student responses.
- Flag responses for review by senior staff.
- Navigate to previously scored responses and amend scores (except for double-scored responses).
- View large tasks with full functionality.
- Enter a training mode to score pre-scored responses and receive feedback on scoring accuracy.

Scoring team leaders could:

- Select, order, and annotate responses for scorer training.
- Review, check-score, and amend scorers' scores.



- Monitor and address flagged responses.

Scoring administrators could :

- Select, order, and annotate responses for scorer training.
- Assign scorers to teams.
- Generate and view reports on inter-scorer reliability.

Coding student descriptions of parental occupations

Each student was presented two pairs of questions in the student questionnaire regarding the occupation of the student's parents. In each pair the students were asked first to provide a job title for the parent, and second to provide a brief description of the work completed by the parent in their job.¹⁰ The responses in each pair (job title and job description) were then coded according to the International Standard Classification of Occupations (ISCO) framework (ILO, 2012).

The AM Marker system was not suitable for conducting the parental occupation coding. Students' text responses to each pair of questions were first exported from the AM Manager system (a feature available to ISC staff) and provided to national centers in a pre-configured MS Excel workbook. The MS Excel workbook included the facility for users to view and code each pair of responses. Once the occupation coding had been completed within each country, the national center then returned the MS Excel workbook to the ISC and the coded data were included in the international database.

3.9 Managing the online systems and data collection

Student instruments (AM Manager)

The ICILS 2023 administration module, AM Manager, was used by national center staff to manage the delivery of the student instruments. It supported the monitoring of test session progress (i.e., submission status), the creation of user accounts, and the assignment of roles for users of related modules (e.g., scoring and translation). It also provided functionality to export data for integration into the IEA WinW3S system.

The AM Manager enabled users to:

- Create user accounts and allocate roles;
- Download the ICILS 2023 test software image (including the test delivery engine);
- View national test session details;
- Monitor the status of national test sessions; and
- Export test session data for import into IEA WinW3S.

Principal, ICT coordinator, and teacher questionnaires delivery (Monitor component)

The Monitor component of the OSS system allowed national centers to track the real-time progress and submission status of the online questionnaires for principals, ICT coordinators, and teachers. Once the questionnaires had been activated for data collection, national centers received access credentials and monitoring rights from the ISC.

The Monitor component enabled users to:

- View a list of respondents who had logged into the system at least once;

¹⁰ The questions presented to students are provided with the [ICILS 2023 IDB](#) and [user guide](#).

- See the most recent question viewed by each respondent;
- Track the overall submission status of each questionnaire; and
- Export participation data in Excel format for further processing.

The exported file included the IDs of principals, ICT coordinators, and teachers, provided that the respondent had submitted at least one valid response.

3.10 Challenges with computer-based delivery in ICILS

Successful collection of data in large-scale CBAs rely on individual participants being able to provide responses on computers in controlled standardized environments from which data can be recorded, organized, and stored for later use. When the data being collected is assessment data, it is imperative that all respondents experience the tasks in an identical manner.

In the CBA environment, variations in the delivery infrastructure can make it difficult to maintain standardization of the test taking experience. This in turn can potentially influence respondents' performance and these differences may be especially pronounced in complex tasks, such as those involving (simulated) multimedia applications or other types of rich or interactive stimulus. These tasks place greater demands on delivery methods than standard, flat, stimulus material, and multiple-choice response options.

The AM test-delivery architecture, including its USB-delivery option, was designed to minimize the potential for variation in students' test-taking experience. The on-screen appearance of the ICILS 2023 instruments could be checked (as a feature of the translation module) throughout the translation and verification processes and was formally verified during the layout verification process. [Chapter 5](#) of this report describes these processes in detail.

The data collected during ICILS 2023 needed to be extracted, transformed, and then loaded into systems for processing, weighting, and analysis. These tasks required the interaction of a complex set of survey operations software applications and systems (as shown in [Table 3.1](#)). As is the case with any system architecture comprising interconnected applications and systems with a range of specific functions, the interfaces between each component are potential points of failure. To ensure that the system worked successfully, study coordinators needed to monitor both the integrity of the functioning of the individual components of the system (paying special attention to ensuring results data were being stored) and the interfaces between the components.

It is not possible in this section of the report to look at the full range of sources and types of information described above. Instead, the focus is on key findings relating to one particular type of information, that is, the proportion and typology of cases that the NRCs and their staff classified as technical problems during administration. Non-response at the unit or item level, whether due to a lack of cooperation, technical issues, or flawed administration, is one of the more serious factors influencing the robustness and stability of inferences from sample studies. Those responsible for implementing a study nationally therefore regularly monitor response and participation rates, especially as these are used as key indicators of success, quality, and fitness for use.

During ICILS 2023, the inspection of unweighted yet cleaned data from all 34 participating education systems (147,637 students from a total of 6,696 sampled classes) yielded some interesting insights. Test administrators were asked to code the participation status of each sampled student on a student tracking form, and the coding scheme included codes for technical problems before, during, or after the CBA sessions; participation in the test and questionnaire sessions were coded separately.

[Table 3.2](#) presents aggregated percentages of these codings for all sampled students (so excluding any non-cooperating schools). Percentages are given across all countries given that the per-country proportions were similar.

Table 3.2: Unweighted participation status percentages across participants based on full database and original coding by test administrators (reference: all sampled students)

Participation status	Test	Questionnaire
Left school permanently		0.9%
Parental permission denied	0.7%	0.9%
Absent	8.4%	8.7%
Incompatible or failed equipment before assessment		1.0%
Technical failure during assessment	0.1%	0.3%
USB stick lost or upload failed after assessment		0.3%
Participated	88.6%	87.8%
Total		100%

Table 3.2 shows that test and questionnaire data were collected from 88.6 percent and 87.8 percent of sampled students respectively. The slightly lower percentage of data collected for the questionnaire is a result of two factors:

1. Students may have been given a break between the test and the questionnaire so some of the students who completed the test may not have returned to complete the questionnaire.
2. In some countries parents gave permission for their children to complete the test but not the questionnaire.

Approximately 9.3 percent of the sampled students had either permanently left school (0.9%) or were absent on the day of testing (8.4%). The slightly higher percentage of students absent during the questionnaire session reflects that some students, for various reasons, did not stay after the first test session.

Please note that the above percentages are based on raw data from the database and consequently may not match other figures relating to participation rates presented in previous chapters and the ICILS 2023 international report. The presented proportions are from unweighted data and as initially assigned by test administrators. We therefore subjected these data to further adjudication. We identified some cases initially coded as a technical problem during the test or questionnaire session, which had (at least partial) responses and should have been classified as participation by the students concerned. These cases, in which a technical problem had been flagged in the student tracking forms but eventually been resolved (as evidenced by the presence of student data) corresponded to incorrect flagging in the respective student tracking forms. It is also possible that a small number of technical problems went unnoticed by the test administrators. Chapter 11 describes how data with some residual technical failures (evidenced or assumed) were eventually treated during the calibration and scaling of test data.

A rather severe issue occurred at the beginning of the data collection for two countries that opted to collect student data via the local server method (see above), both of which met the minimum computer requirements specified by the external software provider to the ISC. However, during the first testing sessions in both countries, it quickly became apparent that the local network server method caused significant delays in loading test content to an unsustainable extent, even though the requirements were met. As a result, both countries successfully switched to the internet (browser-based) method.

One country that administered the student instruments using the online method experienced issues with inaccessible student instruments due to an error from the external provider on a day when a significant number of test sessions were scheduled. The issue was resolved by rescheduling the affected



test sessions. Several other countries were also affected by this outage, but to a lesser extent. Most were able to reschedule testing for the affected schools for another day.

A similar table to [Table 3.2](#) is also available in the ICILS 2018 technical report, allowing for percentage comparisons (Fraillon et al., 2020). Compared to 2018, technical problems before the assessment appear to have increased from 0.1 percent to 1 percent in ICILS 2023, likely due to the previously mentioned issues. However, issues encountered during and after the administration remained consistent. Note that the online administration mode was not available in ICILS 2018.

References

- Fraillon, J., Carstens, R., & Meyer, S. (2020). Computer-based assessment systems. In J. Fraillon, J. Ainley, W. Schulz, T. Friedman, & D. Duckworth (Eds.), *IEA International Computer and Information Literacy Study 2018 technical report* (pp. 31–39). International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>
- ILO. (2012). *International standard classification of occupations: ISCO-08 Volume I*. International Labour Office. <https://www.ilo.org/ilostat-files/ISCO/newdocs-08-2021/ISCO-08/ISCO-08%20EN%20Vol%201.pdf>

Chapter 4:

ICILS questionnaire development

Mojca Rožman and Julian Fraillon

4.1 Introduction

This chapter describes the development process and the content of the ICILS 2023 field trial and main survey questionnaires. The questionnaires collect information to support understanding of the contexts in which students' computer and information literacy (CIL) and computational thinking (CT) learning develops, including the existing conditions, the surround learning, and the processes that contribute to learning. In addition, questionnaire data can contribute to our understanding of variation in students' CIL and CT achievement. Separate questionnaires were completed by students, teachers, principals, and information and communication technology (ICT) coordinators in sampled schools. Another questionnaire, the National Contexts Survey, was designed to collect characteristics of education systems and was completed under the supervision of National Research Coordinators (NRCs), who could obtain the necessary information from relevant national stakeholders or from national documents. Each cycle, the questionnaires are updated to reflect changes in ICT use and new technological developments and to focus on existing or new areas of research interest. In ICILS 2023, examples of new areas of focus were school leadership and technology, approaches to teaching, and the use of generative AI tools in schools. All ICILS questionnaires were administered on a computer.

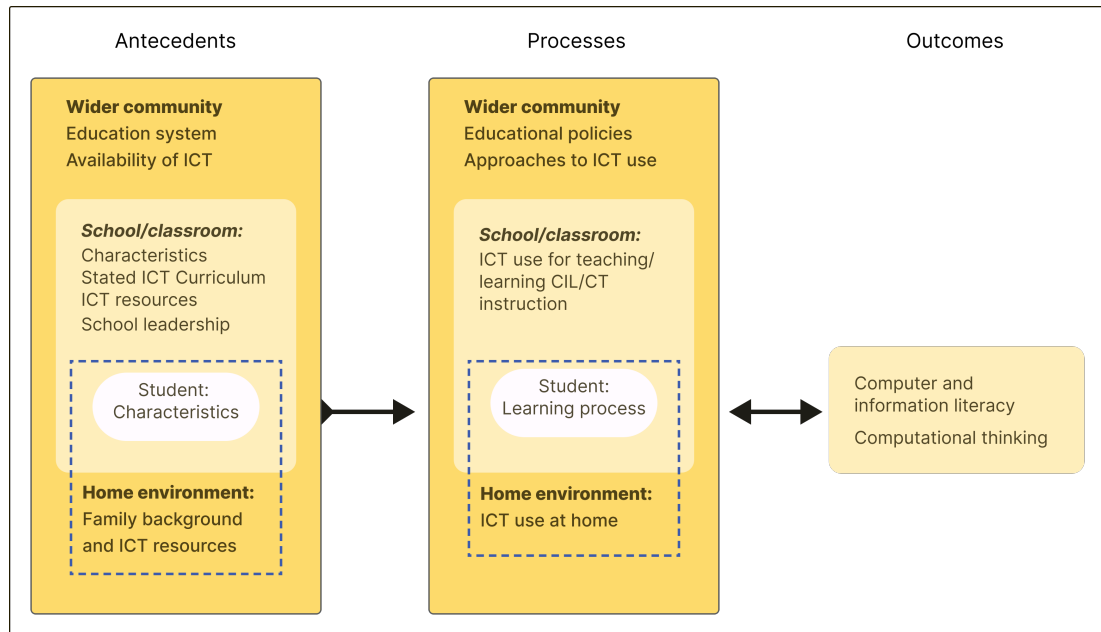
4.2 The conceptual framework and questionnaire content

The contextual framework, which is part of the assessment framework of ICILS 2023 (Rožman et al., 2024), identifies relevant topics related to the environment in which learning of CIL and CT takes place. It assumes that young people develop their understandings of ICT through activities and experiences that take place in schools and classrooms, as well as in their homes and other places outside school.

Since the first cycle in 2013, ICILS has used two main dimensions for the contextual framework. The first dimension is described by the level of the environment in which learning takes place, ranging from the wider community to the level of the individual student, and the second dimension relates to the conditions (antecedents) and activities (processes) for learning (Rožman et al., 2024). The combination of these two dimensions, form the basis of the ICILS context assessment categories. The categories are shown in [Figure 4.1](#).

The ICILS 2023 questionnaires cover all aspects across the different categories of the contextual framework. Information on the wider community is collected through the national context survey, principal, ICT coordinator questionnaires, and other sources. The school-level antecedents are collected through the principal, teacher and ICT coordinator questionnaires, which also include information on processes at school level. The student questionnaire also provides some additional information on school level processes. The student and their home level antecedents and processes are collected through the student questionnaire. Examples of specific topics included for each of the levels and their allocation to the contextual framework categories are presented in [Table 4.1](#).

Figure 4.1: Contexts for ICILS 2023 CIL/CT learning outcomes



Notes: The double arrow between process-related factors and outcomes emphasizes the possibility of a reciprocal association between learning processes and learning outcomes. The single arrow between antecedents and processes indicates the assumption within the ICILS contextual framework of a unidirectional influence between these types of contextual factors.

Table 4.1: Mapping of variables to contextual framework (examples)

Context level	Antecedents	Processes
Wider community	Example antecedents Structure of education Availability of ICT Data sources NCS, PQ, ICQ, and other sources	Example processes Role of ICT in curriculum Approaches to ICT use Data sources NCS, PQ, ICQ, and other sources
School/classroom	Example antecedents School characteristics ICT resources School leadership Data sources PQ, ICQ, and TQ	Example processes ICT use in teaching and learning CIL/CT instruction Data sources PQ, ICQ, TQ, and StQ
Student	Example antecedents Gender Age Data source StQ	Example processes ICT activities Use of ICT CIL/CT Data source StQ
Home environment	Example antecedents Parent socioeconomic status ICT resources Data source StQ	Example processes Learning about ICT at home Data source StQ

Data sources: NCS = national contexts survey; PQ = principal questionnaire; ICQ = ICT coordinator questionnaire; TQ = teacher questionnaire; StQ = student questionnaire.

4.3 Development process

The International Study Center (ISC) at IEA coordinated the development of the ICILS 2023 context questionnaires.

The development of the field trial questionnaires started with a review of the 2018 versions with particular emphasis on the relevance of the preexisting content to the new cycle and the extent of use of previous questionnaire material in reporting or additional analysis. Additional new topics considered relevant to the study were also discussed within the ISC, with NRCs and with the questionnaire development group (QDG). For the field trial, the administered questionnaires included more content than was expected to be used for the main survey. The purpose was to test some of the material and then select the better performing version for the main survey. Two new topics of specific interest identified for inclusion in ICILS 2023 were school leadership for technology and approaches to teaching with and about ICT.

In ICILS, the international version of each questionnaire comprises a common set of questions and items required to be included in the national questionnaires of all participating countries. In addition, where it is deemed to be appropriate, the questionnaires may include a small number of questions and/or items that national centers can choose to include or exclude from their national questionnaires. These optional questions and items are hereafter referred to as *international options*. National centers also have the opportunity to add a small number of questions or items of specific national interest or to add or amend response categories to questions or items (hereafter referred to as *national options*). The use of national options is subject to the approval of the ISC, primarily on the condition that their inclusion does not compromise the collection of data to be included in the international database. Communication between national centers and the ISC regarding national centers' use of international options and inclusion of national options is a key part of the instrument adaptation process presented in detail in [Chapter 5](#).

The development work included development of new material, extensive reviews, and discussions with different stakeholders at different stages of the process. For the first time in ICILS, a QDG was formed to accompany the process of questionnaire development in ICILS 2023. The members of the QDG were: Frederik Aspling (NRC Sweden), Jeppe Bundsgaard (NRC Denmark), Sara Dexter (the United States), Birgit Eikermann (NRC Germany), Julian Fraillon (IEA), Mojca Rožman (IEA), Ronny Scherer (Norway), Anabela Serrão (NRC Portugal), Meng Jung Tsai (NRC Chinese Taipei), and Gary Wong (Hong Kong SAR). The group included experts on digital literacy education from a range of perspectives with familiarity of different national contexts.

The development process for the student, teacher, and school-level questionnaires followed a similar timeline as the CIL and CT test development and can be described in three phases (a detailed timeline is presented in [Table 4.2](#)):

- *Phase 1* (March 2020 to September 2021): previous ICILS questionnaires were reviewed and field trial material was developed with the content being guided by the assessment framework. It also included various rounds of consultations and reviews by the QDG members and NRCs.
- *Phase 2* (October 2021 to September 2022): Preparations for the field trial took place from October 2021 to January 2022. From March to June 2022 the regular international field trial was conducted in 28 participating ICILS countries. Subsequent analyses of field trial data took place from July to September 2022 to inform the suitability of questionnaire material for the main survey. The analysis reports included data from 28 countries for principals and teachers, 27 for ICT coordinators, and 25 for students.
- *Phase 3* (October 2022 to November 2022): The final phase focused on the selection of items for the main survey. Based on field trial analysis results the QDG suggested the materials for the use in the main survey. These were discussed at the NRC meeting in September 2022. Afterwards, based on the feedback from the QDG members and NRCs the final selection of main survey material was prepared by the ISC.

The NRC and QDG meetings were critical points in the test-development process. The first meeting of NRCs coincided with the early phases of the COVID-19 pandemic. As a consequence of the pandemic, this meeting, as well as the second and third NRC meetings were all conducted online rather than in-person. All QDG meetings were conducted online.

Table 4.2: Detailed process of questionnaire development

Year	Month	Group	Activity
2020	March	First meeting of national research coordinators (online)	Reflections on ICILS 2018 questionnaires
2020	26 November – 1 December	Second meeting of national research coordinators and follow-up sessions (online)	Discussion of first ideas for revisions of 2018 questionnaires
2021	9 March	QDG (online)	Planning of QDG work
2021	29 March	QDG (online)	Discussion of new content areas for questionnaires
2021	13 April	QDG (online)	Review of potential questionnaire new material
2021	27-29 April	Check-in meeting with national research coordinators (online)	Discussion of potential new questionnaire material
2021	18 May	QDG (online)	Revision and preparation of questionnaires for further NRC review
2021	June/July	Third meeting of national research coordinators (online)	Review and discussion of final content for field trial questionnaires
2021	27 July	QDG (online)	Review of NRC feedback and discussion of remaining issues finalization of field trial questionnaire content
2022	2 February	QDG (online)	Discussion of NCS content
2022	14 September	Fourth meeting of national research coordinators (in person)	Discussion of final selected content for the main survey questionnaires and rating of priority for inclusion
2022	30 September	QDG (online)	Discussion of NRC meeting outcomes and establishing a proposal for the final selection main survey questionnaire content
2022	October	QDG (online)	Finalization of questionnaire content for the main survey

The procedures and criteria used to review the field trial material and results were the same for the student, teacher, and school-level questionnaires. During instrument development, particular attention was paid to the appropriateness of questionnaire material for the large variety of national contexts and target groups in participating countries as well as to existing differences between education systems



and between schools within each participating education system. The following criteria informed the selection of item material for the main survey:

- Relevance with regards to the ICILS 2023 assessment framework;
- Relevance for the national contexts of the participating countries; and
- Psychometric properties of single items and items designed to measure latent traits.

The analyses of the field trial data included inspections of descriptive statistics at the international and national levels, with special attention to the response frequencies across response categories. Scale statistics were computed for questions in which the items were intended to be used in scales. When considering suggestions for revisions to items used in previous ICILS cycles, the relative merits of the suggested revision to each item were compared to the impact of forgoing the use of data from any trend analyses. These decisions were made on an item-by-item basis.

The National Contexts Survey is administered to the NRCs and does not follow the regular questionnaire development procedures. Its development procedure is described in a separate section below.

Student questionnaire

The student questionnaire was developed to collect information about the students' characteristics and their home background as well as their use of and familiarity with ICT. The proposed content for the ICILS 2023 field trial student questionnaire related to the information of students and their home characteristics was, with a few exceptions, the same as in 2018.

Two changes related to collecting data about the quality of students' access to ICT at home. One new question asked students about the extent to which problems with the internet at home affected their capacity to complete their schoolwork, and the second asked students whether or not they had access to a computer at home to whenever they needed it to complete schoolwork. The *Use of ICT* section was updated as well. The question on ICT use was expanded to distinguish between the use at school for schoolwork, at school for other purposes, outside of school for schoolwork, and outside of school for other purposes. In addition, the extent of use of ICT outside of school was differentiated between school days and non-school days. Two new questions on screen time (students' reports of limits on device use placed on them by their parents) were developed for the field trial: one about the screen time limits and another about screen time enforcement. These two questions were optional at the country level for inclusion in the questionnaire.

The question and items about learning ICT tasks were updated. New items were proposed to measure learning about internet-related tasks and software tasks, at school and outside of school. A new question on learning about responsible ICT use at school was added, and questions about academic-media multitasking and frequency of ICT use for school-related purposes were added as well.

In order to collect data from a larger pool of items than could be reasonably completed by a student in a single session, the field trial student questionnaire was administered in two different forms. The ICILS field trial student questionnaire material included a total of 38 questions (including two international options), with the total of 196 items (151 in form A and 149 in form B). In the field trial, data were collected from 20,034 students from 706 schools in 27 countries.

The final main survey international student questionnaire consisted of 27 questions containing 135 items (including an international option) to be completed within the targeted time of 20 to 25 minutes. Twenty-eight of the items were designed to capture student background information (including ICT experience, devices, and internet connection at home), and 107 (including two international options) were designed to measure students' use of and familiarity with ICT. The main survey student questionnaire consisted of the following five sections:

- *About you:* this section included questions about the student's age, gender, and expected edu-



cation.

- *Your home and your family*: included questions about the country of origin of the students and their parents, the language spoken at home, their parents' occupations and educational backgrounds, the number of books at home, the quality of the internet connection at home, the number of ICT devices at home, and students' access to ICT devices for schoolwork.
- *Your general use of ICT*: included questions about students' experience with ICT, their ICT use in different places and for different purposes, the question about the screen time limits (this was an international option), their learning about various internet related and ICT tasks at and outside of school, and their learning about responsible ICT use at school.
- *Your experience of using ICT to do schoolwork*: included questions about academic-media multi-tasking, frequency of ICT use during lessons for different subjects, and software application use in lessons.
- *Your thoughts about using and learning about ICT*: included questions about self-efficacy, two questions about attitudes towards ICT, a question about learning ICT tasks at school, and one on studying a ICT-related subject in the current school year.

The student questionnaire was administered on a computer. The computer delivery system was configured to support the quality of data provided by students. The system could, for example, prevent students from giving invalid responses (such as selecting multiple contradictory answers where only one was required) and direct students to targeted questions on the basis of a given response. Furthermore, when answering the questions on parental occupation, students were first asked if their parents were currently in a paid job and were then directed to more specific questions about each parent's current occupation (if currently in a paid job) or previous occupation (if not currently in a paid job).

Principal questionnaire

The principal questionnaire was designed to gather information about the school context in general and the schools' systematic approaches to supporting the use of ICT in teaching and learning in particular. The proposed principal questionnaire for the field trial included questions about school-level and principal characteristics. These questions were identical to the content of the ICILS 2018 questionnaire. A new section on school leadership was added, which included 12 newly developed questions. An ICILS 2018 question about procedures to monitor teachers use of ICT for specific learning outcomes was replaced by items related to school leadership and ICT. ICILS 2018 items relating to responsibility for decisions about ICT in the school were replaced by items evaluating distribution of ICT leadership. A question about the principal's age was added, as was a question about attitudes towards the usefulness of ICT to support teachers' work. The field trial questionnaire included 28 questions (including two international options) with 147 items (including two international options). The data analyzed from the field trial questionnaire were provided by 635 school principals from 28 countries.

As the ongoing and future relevance of COVID-19 disruption was unclear at the time of the field trial preparations, we decided to postpone the decision on whether or not to include some questions on this topic in the school questionnaire. After the field trial, the ISC team, in consultation with NRCs and the QDG, decided to include two additional questions associated with the impact of the COVID-19 pandemic on schools (one on the duration of the disruption to teaching and learning and one on the impact of COVID-19 on teaching and learning). In addition, a question on school support for the effective use of ICT in teaching and learning was also added after the field trial.

The selected content of the principal questionnaire for the main survey resulted in a reduction in the size of the principal questionnaire compared to the field trial. The principal questionnaire for the main survey contained 24 questions (including two international options) with 143 items (including two international options) in the following three sections:

- *About your school*: this section contained questions about school size by gender, grades taught



at the school, number of teachers, community size, school management, one question about the school socioeconomic background composition of students (this was an international option), two questions COVID-19, one about disrupted or online instruction because of COVID-19 measures, and one about COVID-19 effects on different aspects of teaching and learning, questions about distribution of ICT leadership, policies regarding aspects of ICT use in the school, importance of ICT related outcomes of education, and expectations of teachers' ICT related knowledge and skills acquisition.

- *Using ICT in teaching and learning at your school:* contained questions about the school vision, a question about peer monitoring in the use of ICT in teaching and learning (this was an international option), shared understanding of ICT in school, frequency of allocated times for teachers to reflect on ICT and pedagogy, priority given to ways of facilitating the use of ICT in teaching and learning, and teachers' participation in professional development.
- *You and your use of ICT:* this section school principals reported their gender, age, and their ICT use for different activities.

Optional generative artificial intelligence (ChatGPT) questions

At the end of 2022, when the ICILS questionnaires were finalized and made available for translation in the ICILS countries, ChatGPT was launched. Based on the rapid and widespread recognition and interest in ChatGPT, and the use of generative artificial intelligence (AI) in schools more broadly, the ISC, in consultation with NRCs and the QDG, the ISC team decided to address the topic of generative AI tools in ICILS 2023. The decision was made to create a set of additional questions as an international option for the principal questionnaire. The decision to limit data collection to school principals was taken primarily to minimize the operational burden on countries. The ChatGPT questionnaire was developed rapidly and separately from the development schedule and processes used to develop the ICILS 2023 questionnaires. The QDG contributed to the initial development of the questionnaire and reviewed the final draft produced by the ISC. National research coordinators were also given a brief period of time to provide feedback on the ChatGPT questions.

The ChatGPT questionnaire included nine questions and 53 items on frequency of ChatGPT use, existence of school policy on ChatGPT use, existence of recommendations on ChatGPT use, whether they had received information from the national authority on ChatGPT use, restrictions on ChatGPT use for teachers and students, frequency of teacher and student use of ChatGPT, reference to ChatGPT use in the curriculum, and three questions on beliefs about consequences of ChatGPT use for teachers and students.

The content of the ChatGPT international option was made available to countries in mid-2023, after the main ICILS data collection had been completed in the Northern Hemisphere countries and before it had started in the Southern Hemisphere countries. In countries choosing to collect data using the international option, it was administered as a separate questionnaire to principals in Northern Hemisphere countries and as part of the questionnaire in Southern Hemisphere countries. The ChatGPT international option was completed by principals in 12 ICILS countries.

ICT coordinator questionnaire

The ICT coordinator questionnaire was designed to collect information about the ICT resources in schools and the support for the use of ICT in teaching and learning. The ICILS 2018 ICT coordinator questionnaire was revised to include new questions: one about the role of the ICT coordinator in the school, a new section on school leadership with eight questions, new questions about the provision of software for smartboards by the relevant education authority, as well as three items about monitoring software (these three were an international option). As the questions in the ICT questionnaire covered many different topics, we also included a question about how many people needed to be consulted to complete the questionnaire. The field trial questionnaire contained 24 questions (three of which comprised the international option) with 141 items (three of which comprised an international option). The data analyzed from the field trial were provided by 634 ICT coordinators from 27 countries.

The final ICT coordinator questionnaire comprised 19 questions (including the three questions comprising the international option) with a total of 119 items (including five international options). It contained the following four sections:

- *About your position*: this section included two questions about the ICT coordinators position at school, their duties and their school's experience with computers for teaching and learning.
- *ICT technical resourcing in your school*: included questions on the availability of technology infrastructure, software, and hardware resources at school, number of device types, computers connected to the internet in the school, number of smartboards and provision of software for smartboards, provision of ICT devices for teachers, percentage of students with access to portable computers, location of ICT devices, and three questions about monitoring software, including its use (these three were an international option).
- *ICT support in your school*: this section consisted of questions on the technical and pedagogical support provided for ICT use at school and the extent to which a lack of resources was hindering that use.
- *Using ICT in teaching and learning at your school*: contained questions about the school vision for using ICT, availability of learning materials for ICT teaching and learning, shared understanding of the use of ICT in the school, about the source to assess effectiveness of the professional learning activities for ICT use in teaching and learning, and pedagogical resourcing hindrances to the use of ICT in teaching.

Teacher questionnaire

The teacher questionnaire was designed to collect contextual information about the school and classroom contexts for ICT learning, the use of ICT for teaching and learning, teachers' views on the pedagogical use of ICT, and their confidence in using computers. The teacher questionnaire for the field trial included questions on teacher characteristics, initial education, and professional learning activities. With the exception of initial education, where items were heavily edited for clarification, this content remained unchanged from 2018. Similar to the principal and ICT questionnaires, eight questions on school leadership for ICT and a question on whether ICT hinders or helps in the work as a teacher were added for inclusion in 2023. A question on the frequency of ICT-based activities was added, as well as two questions on teaching approaches: a question on the allocation of time to different teaching activities, together with the extent of ICT use in these activities, and a question on teaching beliefs. The question on the allocation of time for different classroom activities was tested in two different versions, the first asking for the approximate percentage of time that students spent on different activities and the second asking teachers to choose from the response options given how often students spent on different activities. The question about teaching beliefs was also trialed using two alternative versions, one with the traditional Likert-type response options and the other requiring teachers to rank the degree to which statements reflected their beliefs.

As with the student questionnaire, our intention was to reduce the administration time for teachers, and the teacher questionnaire was administered in two different forms. The teacher questionnaire for the ICILS field trial contained a total of 30 questions and 286 items. Form A contained 22 questions and 151 items and Form B contained 22 questions and 152 items. The analyzed data from the field trial questionnaire included 8,323 teachers from 710 schools in 28 countries.

The final teacher questionnaire comprised 21 questions (including three optional questions) with a total of 192 items (including 45 optional items). It contained the following five sections:

- *About you*: this section contained questions about the teacher's gender, age, subjects taught, and if they teach in the target grade.
- *Your use of ICT*: contained questions about experience with ICT for teaching purposes, general use of ICT, and a question related to their self-efficacy in ICT use.



- *Learning to use ICT in teaching*: asked teachers about whether their initial teacher education included learning to use ICT, whether they had participated in ICT-related professional learning, and their use of ICT in teaching and learning at school (a question about the use of software tools in class, and a question about emphasis on computational-thinking skills in teaching were international options).
- *Using ICT in teaching and learning at your school*: included a question about the existence of a school vision for using ICT, whether there is a shared understanding of the ICT use at school among teachers, their perceptions on school ICT availability, and perceptions of positive and negative outcomes of ICT for students.
- *Teaching and learning with ICT*: these questions asked teachers to select a reference class, provide information about the subject taught in that class, how often they use various teaching activities, and state whether they used ICT for teaching and learning activities in this class. Further teachers had to indicate the proportion of lessons in which they use specific ICT based activities, the frequency of their use of software tools, the emphasis they give to specific ICT based and CT capabilities. The final question asked about their teaching beliefs (this question was an international option).

National Contexts Survey

The development of students' CIL and CT is embedded in their national contexts and should be interpreted with consideration of these national characteristics. The national contexts survey (NCS) provides an opportunity to collect information at the national level on the general characteristics of the education system as well as policies on, and the curricular background of, CIL and CT education. The national center in each participating country was responsible for the completion of the NCS. Given the broad nature of the content and the expectation that the responses provide defensible objective data relating to each country, national center staff were encouraged to work, as necessary, with others in their countries (e.g., ministries and departments of education, relevant non-government organizations, specialist organizations concerned with supporting the application of educational technologies, and teachers' associations) to complete the survey. The NCS in ICILS 2018 was influenced by the experience and outcomes of studies such as the Second Information Technology in Education Study 2006 (Plomp et al., 2009), the U.S. Department of Education, Office of Educational Technology (2011) study of educational technology, and ICILS 2013 (Fraillon et al., 2014). The NCS in the current ICILS cycle was derived from the NCS developed for use in ICILS 2018.

The content to be included in the NCS was developed with consideration of the planned content and structure of the ICILS 2023 international report chapter on *National and school contexts for computer and information literacy and computational thinking education* (see Fraillon, 2025, Chapter 2). When planning both the report chapter and the NCS, the ISC, together with NRCs and the QDG, reflected on the content of the equivalent chapter in the ICILS 2018 international report. This chapter included a brief profile of each country's education system and approach to providing education for ICT. These country profiles were collected separately from and after the 2018 NCS data collection. As the number of participating countries increased significantly compared to the previous cycle, and this would also increase the length of the corresponding chapter in the international report, the inclusion of country profiles was discussed with NRCs and the QDG. There was general agreement among the NRCs, the QDG, and the ISC staff that brief descriptions of each participating country would provide a useful context for the results and should again be part of the international report. Specifications for the content and length of the country profiles, as well as the facility for national centers to submit their country profiles, were included as part of the NCS.

Further discussion included a close examination of the 2018 NCS content to determine the relevant content for ICILS 2023. A focus was placed on the content reported in 2018, and items were revised to include more information on the specifics of CT education. In addition, we focused on the relevance of the content in relation to the ICILS 2023 assessment framework, the relevance and added value of collecting information about the wider community context of CIL and CT education, and the



appropriateness of the question for the national contexts of the participating countries. The ISC team and the QDG reviewed the content and presented a draft version, which was reviewed by the NRCs. Once the final version was agreed, the questionnaire was implemented as an online questionnaire.

The final version of the NCS consisted of 19 questions with 122 items (some items were closed responses, and others asked for free text responses); most items also had an option for respondents to add additional comments if they felt they were warranted. The questions also addressed content associated with key antecedents and processes in relation to CIL and CT education in each country. The questions were grouped into two sections:

- *The education system*: this section started with the descriptive country profile and further included questions on some general characteristics of the country's education system, questions about how CIL and CT are taught at different International Standard Classification of Education (ISCED) levels, and questions about how much autonomy schools have in various aspects related to CIL and CT teaching and learning.
- *Approaches to CIL/CT education in ICILS countries*: questions in this section were related to general policies and overall approaches to ICT in school education in the countries, assessment of CIL/CT, development of teacher capacity to use ICT, teacher rather than specifically to the target grade

The online questionnaire delivery platform used for the NCS enabled national center staff to log on and off in order to complete the questionnaire in sections over time rather than in one session. This enabled national center staff to research responses to the questions before entering them into the system and review and revise their responses before submitting the completed instrument. National centers were encouraged to conduct internal reviews of the information they provided in the international report and to seek advice from local experts where appropriate. The ISC used the outcomes of the national contexts survey in conjunction with data from externally published sources to inform the descriptions of the education systems participating in ICILS that were included in the ICILS 2023 international report.

References

- Fraillon, J. (Ed.). (2025). *An international perspective on digital literacy: Results from ICILS 2023*. Springer. <https://doi.org/10.1007/978-3-031-87722-3>
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA international computer and information literacy study international report*. Springer Nature. <https://doi.org/10.1007/978-3-319-14222-7>
- Plomp, T., Anderson, R., Law, N., & Quale, A. (2009). *Cross-national policies and practices on information and communication technology in education* (2nd). Information Age Publishing.
- Rožman, M., Fraillon, J., Dexter, S., Bundsgaard, J., & Schulz, W. (2024). Contextual framework. In J. Fraillon & M. Rožman (Eds.), *IEA International Computer and Information Literacy Study 2023 assessment framework* (pp. 45–55). Springer. https://doi.org/10.1007/978-3-031-61194-0_4
- U.S. Department of Education, Office of Educational Technology. (2011). *International experiences with technology in education: Final report* (M. Bakia, R. Murphy, K. Anderson, & G. E. Trinidad, Eds.; tech. rep.). U.S. Department of Education Office of Educational Technology. www2.ed.gov/about/offices/list/os/technology/iete-full-report.doc

Chapter 5:

Preparation and verification of the ICILS 2023 study instruments

Marta Moreno Hidalgo and Tim Daniel

Introduction

This chapter describes the procedures for preparing national versions of the ICILS 2023 study instruments. The primary objective of the instrument preparation process is to maintain comparability of the national instruments with the international versions, ensure equivalence across countries and languages, guarantee the appropriateness of the instruments for each participant's national context and education system and ensure the consistency of translations across cycles for countries participating in earlier ICILS cycles. To achieve the objective of cross-national comparability and data validity, all national ICILS 2023 study instruments underwent a thorough and standardized preparation process as described in this chapter.

The chapter focuses on the following topics:

- Description of the ICILS 2023 materials that required translation and adaptation, the languages administered, and the systems used for producing national instruments;
- Adaptation, translation, and layout review of national language versions of the international source materials and associated support materials; and
- International adaptation, translation, and layout verification of the national instruments.

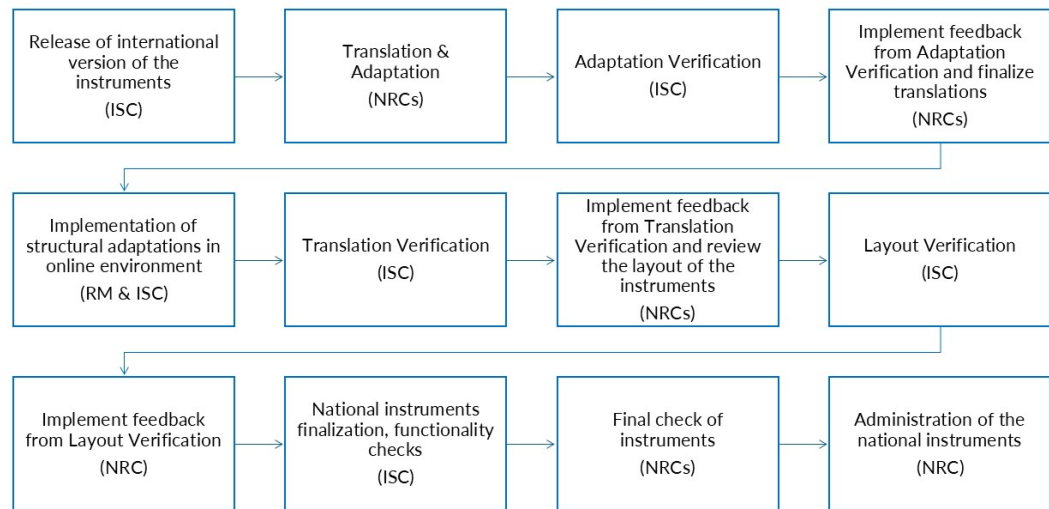
The ICILS 2023 International Study Center (ISC) at IEA developed the international versions of the student assessment materials and context questionnaires (in English language) in close collaboration with the national study centers.¹¹

Upon the release of the international version of the instruments, all participating countries were required to adapt and, if applicable, translate the international version of the instruments to their national context and language(s). Guidelines for these procedures were provided in the Survey Operations Procedures Unit 3, as well as in supplementary manuals for using the computer-based administration systems. After completing each step of national instrument preparation, the adaptations, translations, and layout of the instruments underwent verification.

The ISC oversaw the instrument preparation and verification procedures. [Figure 5.1](#) shows the overall workflow for preparing the national study instruments. The process was repeated twice—once before the field trial and then again prior to the main survey. To the extent possible, participating countries were encouraged to use field trial translations as a basis for their translations of the main survey instruments. The ISC also managed and coordinated the translation verification process in collaboration with cApStAn Linguistic Quality Control, a leading provider for translation and translation verification services based in Brussels, Belgium.

¹¹ Please note that throughout this chapter we will use the term “countries” when referring to the 34 countries and one benchmark that participated in the study.

Figure 5.1: ICILS 2023 national instrument preparation workflow



5.1 Description of the materials, languages, and systems

ICILS 2023 instruments

The ISC provided National Research Coordinators (NRCs) with the international versions of the study instruments for adaptation and translation (see Table 5.1). The student test modules and questionnaire were adapted, translated, finalized, and administered within RM Results' Assessment Master (AM).

Table 5.1: Instruments that required translation/adaptation

Instruments that required translation/adaptation	Translation mode			Data collection online system		
Context questionnaires						
ICT Coordinator Questionnaire	Microsoft Word documents			IEA's Online Survey System		
Teacher Questionnaire	Microsoft Word documents			IEA's Online Survey System		
Principal Questionnaire	Microsoft Word documents			IEA's Online Survey System		
Student Questionnaire	RM	Assessment platform	Master	RM	Assessment platform	Master
Assessment instruments						
7 CIL modules*	RM	Assessment platform	Master	RM	Assessment platform	Master
4 CT modules**	RM	Assessment platform	Master	RM	Assessment platform	Master

Note: * Administered by all countries

** Optional, administered by some countries

The principal, ICT coordinator, and teacher questionnaires were adapted and translated in MS Word documents before being transferred to the IEA Online Survey System (OSS) for layout verification and subsequent administration. For readability, this chapter refers to the student test modules and questionnaire as "student instruments," while the principal, ICT coordinator, and teacher questionnaires



will be collectively referred to as “adult questionnaires.”

As for previous ICILS study cycles, all countries administered the computer and information literacy (CIL) test. As in ICILS 2018, the computational thinking (CT) test was included as an optional component implemented by countries who decided to participate (for further details on the content of these modules, see [Chapter 2](#)).

National study centers also had to adapt and translate procedural manuals and scoring guides for constructed-response test items. The ISC did not verify the adaptations and translations of these procedural manuals. However, the translations of *Test Administrator Manual* and *School Coordinator Manual* underwent a final quality control spot check by an independent quality observer (for further details on the content of these modules, see [Chapter 9](#)).

Administered languages

In most countries, the ICILS 2023 instruments were adapted and translated into the language(s) of instruction used in the national education system. A country’s decision to administer assessment instruments and questionnaires in multiple languages depended on factors such as national requirements, the sampling frame, and whether schools with instruction in a particular language were included in the sample. Countries could also choose to administer only specific materials in a given language (e.g., only administering the student assessment and student questionnaire and not the adult questionnaires).

For ICILS 2023 a total of 34 countries and one benchmarking participant prepared their national study instruments in 50 languages (see [Table 5.2](#)). Of the participating entities, 10 administered the instruments in more than one language.

Translation system

The student instruments were adapted and translated via the Designer module in AM. The system enabled users to edit translations, maintain a complete record of the editing history, and retrace changes made by other users, ensuring transparency in the translation and verification process. Additionally, the system displayed differences between previously saved translations and the current version in the form of tracked changes. Users could preview the translations as they would appear to students, facilitating the identification of potential layout issues. It also allowed for translation exports in CSV, XLIFF, or JSON formats.

Adaptations requiring changes to the structure of the student questionnaire—such as adding national questions, splitting or omitting answer options (see [Subsection 5.2](#) for more information)—involved a more complex workflow that could only be implemented by the ISC with assistance from RM.

The adaptation and translation of the adult questionnaires, along with their verification processes, were conducted in MS Word files. The countries replaced the English version for their translations and submitted these files for translation verification. After verification, and after implementing the verifier’s feedback within the files, the NRCs transferred the final translations to the OSS and included the approved structural adaptations within the platform. The ISC then verified the accuracy of these adaptations as part of the final layout check before the materials were prepared for administration.

5.2 Adaptation, translation, and layout reviews

Adaptation process

NRCs were required to adapt and translate the international source version of the ICILS study instruments in accordance with the guidelines outlined in Survey Operation Procedures Unit 3, which detailed the procedures and necessary steps.

Adaptations were categorized into two types. Non-structural adaptations referred to linguistic changes that were either required and predefined by the ISC—such as the names of people, places, or terms related to education systems within countries—or requested by the country to address specific na-



Table 5.2: ICILS 2023 administered languages

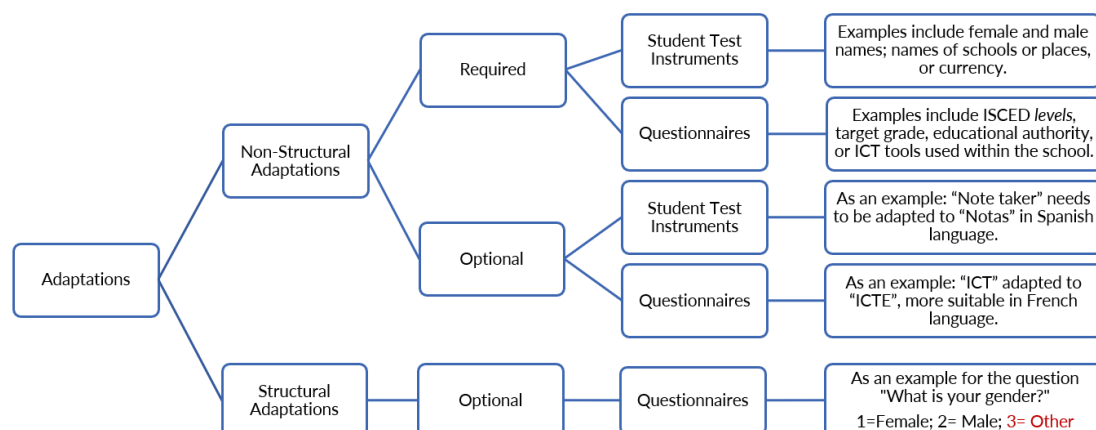
Country	Language	Modules		Questionnaires			
		CIL	CT	Student	Principle	Teacher	ICT Coordinator
Austria	German	•	•	•	•	•	•
Azerbaijan	Azerbaijani	•		•	•	•	•
	Russian	•		•	•	•	•
Belgium (Flemish)	Dutch	•	•	•	•	•	•
Bosnia and Herzegovina	Bosnian	•		•	•	•	•
	Croatian	•		•	•	•	•
	Serbian	•		•	•	•	•
Chile	Spanish	•		•	•	•	•
Chinese Taipei	Chinese (Mandarin)	•	•	•	•	•	•
Croatia	Croatian	•	•	•	•	•	•
Cyprus	English	•		•	•	•	•
	Greek	•		•	•	•	•
Czech Republic	Czech	•	•	•	•	•	•
Denmark	Danish	•	•	•	•	•	•
Finland	Finnish	•	•	•	•	•	•
	Swedish	•	•	•	•	•	•
France	French	•	•	•	•	•	•
Germany and NRW	German	•	•	•	•	•	•
Greece	Greek	•		•	•	•	•
Hungary	Hungarian	•		•	•	•	•
Italy	Italian	•	•	•	•	•	•
Kazakhstan	Kazakh	•		•	•	•	•
	Russian	•		•	•	•	•
Korea, Republic of	Korean	•	•	•	•	•	•
Kosovo	Albanian	•		•	•	•	•
Latvia	Latvian	•	•	•	•	•	•
	Russian	•	•	•			
Luxembourg	German	•	•	•	•	•	•
	English	•	•	•	•	•	•
	French	•	•	•	•	•	•
Malta	English	•	•	•	•	•	•
Norway	Bokmål	•	•	•	•	•	•
	Nynorsk	•	•	•	•	•	•
Oman	Arabic	•		•	•	•	•
	English	•		•	•	•	•
Portugal	Portuguese	•	•	•	•	•	•
Romania	Romanian	•		•	•	•	•
Serbia	Serbian	•	•	•	•	•	•
Slovak Republic	Hungarian	•	•	•			
	Slovak	•	•	•	•	•	•
Slovenia	Slovenian	•	•	•	•	•	•
Spain	Basque	•		•	•	•	•
	Catalan	•		•	•	•	•
	Galician	•		•	•	•	•
	Spanish	•		•	•	•	•
	Valencian	•		•	•	•	•
Sweden	Swedish	•	•	•	•	•	•
Netherlands	Dutch	•	•	•	•	•	•
United States	English	•	•	•	•	•	•
Uruguay	Spanish	•	•	•	•	•	•



tional contexts. Structural adaptations involved modifications to the structure of the questionnaire or individual questions, such as adding, splitting, or removing response categories; introducing national dimensions; including national questions; or omitting questions or dimensions.

The types of adaptations that could be implemented by the NRCs are presented in [Figure 5.2](#).

Figure 5.2: ICILS 2023 adaptation types



The international version of the study instruments indicated where non-structural adaptations were necessary by using square brackets and red font to note terms to be adapted. For example, [country of test] was meant to be replaced with the name of the NRC's country and [target grade] with the name of the target grade (e.g., "Grade 8").

To help participating countries to adapt their materials to the national context, the ISC provided them with notes that included brief descriptions of the required adaptation clarifying the meaning of the terms used (and thus enabling the translators to select the appropriate national term or expression necessary to convey the intended meaning). In addition, general guidelines for translation and adaptation were provided (listed in [Subsection 5.2](#)).

Both required and country-specific adaptations had to be documented by each country in the National Adaptation Form (NAF) and were subject to adaptation verification by the ISC.

National adaptation form

The NAF was an Excel document containing a comprehensive adaptation and verification history of each set of national instruments. Two different NAFs were used for ICILS 2023, one for the student instruments (test and questionnaire) and another for the adult questionnaires. This was to reflect that the instruments were administered on two different software platforms, and that the preparation of the national instruments followed different timelines.

The student instruments NAF consisted of individual worksheets for the CIL modules, the CT Modules (where applicable) and the student questionnaire. The adult questionnaire NAF consisted of individual worksheets for the adult questionnaires as well as a section to report layout verification comments.

Overall, documenting an adaptation in the NAFs required entering the proposed adaptation in the target language, an English back translation of the adaptation, recoding instructions (if applicable and only for structural adaptations to the adult questionnaires), and a justification for any further alterations. For the ease of use and documentation of the different stages of verification, the NAF included designated areas for each item, user, and instrument.

Both NAFs also contained separate worksheets for completion instructions and a style guide. The style guide collected information about:



- The language variant (e.g., British or American English, Maghrebi Arabic or Mashriqi Arabic);
- Language-specific rules (e.g., capitalization rules for languages based on Latin or Cyrillic alphabet, use of diacritic in Russian language);
- Use of anglicisms for technical terms (e.g., using Copy/Paste in English rather than the national language);
- Form of address to the target audience (formal or informal);
- National conventions for numbers and time notations (e.g., use of dot, space, comma, or apostrophe, for thousand separators);
- Grammar and measurement units used in the materials (e.g., use of symbols or full words); and
- An open section for countries to specify other aspects to be considered during the translation work.

At specific points during the instrument preparation process, sections of the NAF were completed and reviewed by NRCs and verifiers. The first version of the NAF was filled out by the NRC in collaboration with national translator(s) and reviewer(s). This occurred while the materials were being adapted and translated. This version was submitted to the ISC, for adaptation verification (i.e., review and approval). The ISC comments were then returned to the NRC for review and to implement any suggested changes. Once all adaptations were agreed upon by the NRC and the adaptation verifiers, the NAF was provided to the international translation verifiers during translation verification so they could check the correct implementation of the adaptations. When translation verification was complete, the NAF was again returned along with the materials for NRCs to review and implement any final changes.

For the adult questionnaires NAF, an additional version was submitted to the ISC after translation verification to document any layout issues identified following the transfer to the OSS. Once these issues were corrected, the ISC returned the final version of the NAF with its comments to the country.

The NAF was an important record of each country's final instrument versions, as it contained information used throughout the different stages of the instrument preparation process.

Translation process

Along with adaptation, another key part of the national instrument preparation process was the translation of the materials into the target language(s).

Countries were strongly advised to use a double translation and reconciliation approach. This method involved two independent translators working on the same materials, followed by a third individual who reviewed and reconciled the translations to select the best final version. When time or financial constraints made this approach unfeasible, a single translation with review method was recommended. In this case, one individual translated the materials, and a second person reviewed the translations. Regardless of the translation procedure used, the NRC was responsible for reviewing and finalizing the translations of the ICILS 2023 instruments.

Essential qualifications for translators and reviewers included:

- Having the target language as mother tongue;
- Excellent knowledge of English;
- Good level of computer literacy;
- Familiarity with survey translations;
- Familiarity with the subject area; and
- Familiarity with education in the target grade.



NRCs received specific guidelines for the translation of the student instruments. These guidelines included:

- Minimizing cultural and contextual differences.
- Considering the use of anglicisms where appropriate for the target population, reflecting the predominance of English-language terminology in computing and internet culture.
- Maintaining the register and meaning between the source and target versions.
- Accurately translating idiomatic expressions, ensuring correct grammar, and ensuring appropriateness for the target population.

Student test instruments trend comparison

A total of six modules for the student test instruments (four CIL two CT) were carried over from the previous ICILS cycle. These modules, known as “trend modules,” were administered across all countries participating in ICILS 2023 (as applicable for the optional CT modules). They were used to measure changes in achievement for those countries that participated in both ICILS 2018 and ICILS 2023. To ensure the quality of comparisons over time, countries that participated in the 2018 cycle were asked to maintain, to the extent possible, the translations they used earlier.

The list of countries and modules administered in ICILS 2018 are listed in [Table 5.3](#). Please note that some countries that participated in the CT module for 2023 did not participate in this option for the 2018 cycle.

A total of two countries participated in ICILS 2013 and ICILS 2023 but not in 2018 (Slovak Republic and Slovenia). Please note that since these countries did not participate in the 2018 cycle, trend was considered as lost for these cases.

Table 5.3: Trend countries (administration in 2018)

Country	Language	4 Trend CIL Modules	2 Trend CT Modules
Chile	Spanish	•	
Denmark	Danish	•	•
Finland	Finnish	•	•
	Swedish	•	•
France	French	•	•
Germany and NRW	German	•	•
Italy	Italian	•	
Kazakhstan	Kazakh	•	
	Russian	•	
Korea, Republic of	Korean	•	•
Luxembourg	German	•	•
	English	•	•
	French	•	•
Portugal	Portuguese	•	•
Uruguay	Spanish	•	
United States	English	•	•



For countries that participated in ICILS 2018, the existing translations of the modules from the previous cycle were made available within the AM Designer upon the release of the international version of the materials. Countries were responsible for reviewing the trend translations and, only when absolutely necessary, requesting and implementing changes to these (e.g., update of a national adaptation or linguistic correction of a translation due to technical errors or mistranslations in 2018). Requested changes to trend translations were required to be carefully documented in the Trend Verification Form (TVF) and were reviewed by the ISC during translation verification and considered during the data adjudication.

Trend verification form

The trend verification form (TVF) was an Excel file with a similar structure to the NAF. The TVF was created to keep track of changes in the translations from the previous to the current study cycle for the student test instruments (4 CIL modules and 2 CT modules from ICILS 2018) and therefore only needed to be completed by countries that participated in ICILS 2018 for the CIL and the optional CT modules.

The TVF consisted of individual worksheets for the CIL modules and CT modules (where applicable) and a separate worksheet with completion instructions. NRCs were instructed to document any requested changes to translations from the materials administered in 2018. To request a change, NRCs needed to provide the translation as it was in 2018, the suggested new translation for ICILS 2023, the English back translation of the proposed change, the location of the translation, the type of change (outdated national adaptation, outright mistranslation, missing text, translation not being correctly copied over or misspelling) and the justification of the change in separate columns of the TVF.

During translation verification, the international translation verifiers reviewed the TVF and provided feedback on whether the proposed change to translation was necessary or not. In addition, the verifiers documented any additional changes to the trend materials not listed in the TVF. Based on the verifier's feedback, the ISC advised the NRC on whether or not the suggested change should be implemented. The final decision to keep or reject the change to trend translation was made by NRCs, considering the comments provided by the ISC and verifiers during the translation verification process.

Layout preparations

The final step in the country's preparation of national instruments was reviewing the layout of the materials. NRCs were asked to review both the student instruments (within the AM Designer) and the adult questionnaires (within OSS) to ensure that the content was accurate, the layout closely matched the international version, and the functionality worked as intended.

For all instruments, NRCs were asked to check the following:

- The content was accurate and matched the translations entered.
- No adaptation brackets or HTML tags were visible.
- National extensions were displayed appropriately.
- All content was displayed correctly (e.g., text did not overflow outside boxes or pages).
- Functionality was working as expected.

For the review of the adult questionnaires, NRCs were asked to transfer their final translations to OSS and document any layout related issues within the layout verification tab of the adult questionnaire NAF. Upon completion of this step, NRCs were required to submit the completed NAF to the ISC together with the final version of the Word and OSS files.

For the review of the student instruments, NRCs received access to a URL that contained all assembled modules to support the layout review of the materials. Additionally, NRCs could make use of the full



and live previews available within the AM Designer module.

While reviewing the layout of the student instruments, NRCs were required to flag, describe and categorize within the AM Designer any layout verification issues by entering a comment linked to the affected text element.

Countries used the following categories to report layout issues:

- **Missing/Unreadable Text:** Examples include elements completely or partly missing on the screen or unreadable due to, for example, an overlap with other text elements.
- **Text Out-of-Bounds:** Examples include text not placed within boundaries.
- **Misplaced Element:** Such as text or images not appearing where they should, compared to the international source version.
- **Other:** Used to report any other type of issue not covered by the previous categories or in case of a combination of categories. For these cases provision of a detailed and clear description of the issue was required from the country.

Once the comments were entered in the system, countries set the status for all student instrument materials to “Ready for LV” status and informed the ISC that the materials were ready for review.

5.3 International verification processes

After the international instruments were adapted, translated, and internally reviewed by the countries, the national versions of the instruments were submitted for international verification, which consisted of a rigorous three-part verification process: (1) adaptation verification, (2) translation verification, and (3) layout verification.

Adaptation verification

NRCs were required to consult with ISC staff for a review of all proposed national adaptations. The ISC particularly emphasized the need for NRCs to discuss and document any adaptation that might result in a serious deviation from the international instruments, as well as providing recoding instructions for structurally adapted questions. Properly documented and verified recoding instructions were of utmost importance, particularly during data processing, where they were implemented to ensure international comparability of nationally adapted questions or question components. This step required NRCs to map data collected in structurally adapted national items to the international items, with their correspondent internationally comparable value schemes. This information was then used in analysis and reporting.

National centers began completing the student instrument and adult questionnaire NAFs after reviewing the international version of the instruments. They then submitted the NAF for review by the ISC. Following its review, the ISC provided the national centers with feedback on their requested adaptations and, where appropriate, made suggestions for better alignment to the international version. Discussions regarding the adaptation and final status were reported across different columns of the NAF.

Common issues identified during review of adaptations included:

- Inconsistent use of adaptations within or across modules and questionnaires;
- Fictional names for software or technologies not considered to be equivalent to the source version; and
- Difficulties in establishing country-appropriate adaptations for ISCED levels.

The student instrument and adult questionnaire NAFs were finalized only after the ISC and national



study center reached an agreement on all adaptations. Once finalized, any updates were incorporated into a new version of the NAF, which was subsequently provided for translation verification.

Translation verification

The national versions of the ICILS 2023 instruments underwent international translation verification. During this step the ISC and cApStAn coordinated in the processing of the materials.

Whereas the ISC was responsible for the training of international translation verifiers, initial review of the country translations, processing of the translation verifier's feedback and returning the materials back to the country, cApStAn was responsible for the selection, in depth training of the translation verifiers, processing of the verifier materials (before being sent to the verifier or returned to the ISC), and overall coordination and communication of the verification processes with the ISC.

The international translation verifiers were responsible for reviewing and documenting the quality and comparability of the national instruments with the international source version of the instruments.

Verifiers were required to be professional translators or interpreters. In addition, they were expected to meet the following qualifications:

- Native speakers of the target language;
- Fluency in English;
- Familiar with the subject area (if possible);
- Located in the target country or in close contact with it;
- A good level of computer literacy; and
- Trained by the ISC and cApStAn.

Verifiers' were trained during an online webinar, where they received detailed information about the study and study instruments, their tasks and role, the verification steps, the materials to be used during verification (NAF, TVF, MS Word Files) as well as the use of the AM Designer for the review of the student instruments. In addition to this, verifiers received access to relevant manuals and instruments, as well as a comprehensive set of instructions and relevant examples for reviewing the national instruments and documenting deviations from the international source version. The ISC and cApStAn also ensured all international translation verifiers received continuous support and provided the verifiers with constructive ongoing feedback as needed.

There were three modes of translation verification in the ICILS 2023 main survey: full verification of new translations, focused verification of translations which were already verified for the field trial or done for English speaking countries, and trend verification of trend translations (only applicable to the CIL/CT modules).

The international translation verification process included the following:

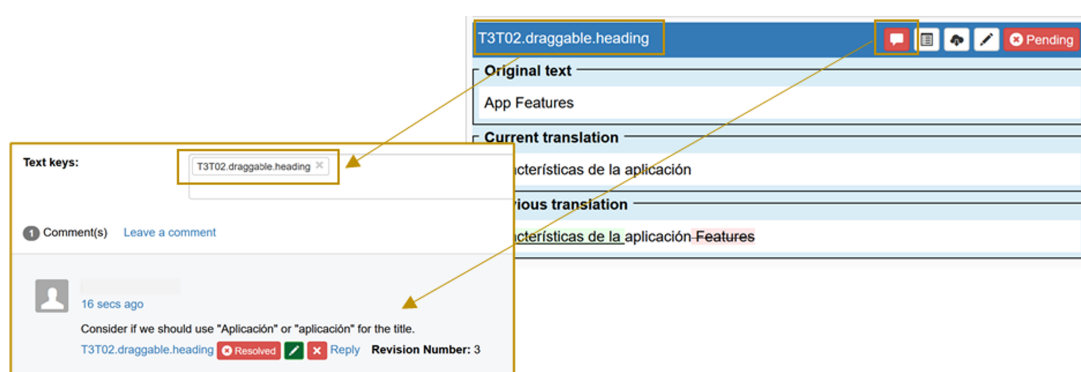
- Checking the accuracy, linguistic correctness, and comparability of the translations and adaptations of the student instruments and adult questionnaires. This included verifying the correct implementation of national adaptations, ensuring the consistent use of words and terms listed in the style guide, and confirming alignment with previously used translations (where applicable);
- Documenting any deviations between the national and international source versions, or between new and previously used translations, such as additions, deletions, and mistranslations; and
- Proposing alternative translations or adaptations to enhance the accuracy and comparability of the national instruments.

Translation verifiers provided their feedback within the AM Designer (for student instruments), the MS Word documents (adult questionnaires) and the NAFs.

For the adult questionnaires, verifiers were asked to correct the translations and/or to add comments specifying errors using the edit and comment functions within the Word File. Changes were implemented in track changes to facilitate the NRC review.

For the student instruments, verifiers were asked to correct the translations and/or to add comments specifying errors using the edit and comment functions within the AM Designer. Corrections or suggestions could be applied directly to the text elements in the online environment. Any changes made by the verifier were displayed as track changes in the text element. Verifiers were required to document their corrections and suggestions by using the comment function for every affected text element to describe the issue (see [Figure 5.3](#)).

Figure 5.3: Track changes after translation verifier edits



For the review of national adaptations and to report severe translation issues, verifiers were required to enter comments in the NAF. Verifiers also used a style guide included in the NAF to ensure that specific country linguistic requirements were correctly implemented throughout.

For student test instruments undergoing trend verification, verifiers focused on checking the identicalness between 2018 and 2023 translations; checking the adequacy and correct implementation of proposed changes to trend translations documented in the TVF; and recording any undocumented differences in trend translations.

In order to document the comparability between the translated materials and the international version, verifiers used a range of severity codes to indicate the nature of any issues that they may have found as well as their severity. The verifiers used the following codes:

- **Code 1:** Major change or error. Examples include incorrect translations that change the meaning or difficulty of the text, additional or omitted information, and undocumented changes to trend translations. When in doubt over the severity of an error or when the verifier was unsure of how to correct a possible error, **Code 1?** was used.
- **Code 2:** Minor change or error. Examples include spelling mistakes and punctuation errors that do not affect comprehension.
- **Code 3:** Suggestion for an alternative. The translation may be adequate, but the verifier suggests a different wording.
- **Code OK:** Acceptable adaptation. The adaptation (either required or country-specific) was acceptable and appropriate.

After translation verification, the ISC reviewed the verifiers' feedback and returned the materials to the



national centers, namely the updated versions of the NAF, TVF, and MS Word files (for the adult questionnaires). Changes to the student instruments within the AM Designer were available to countries immediately after translation verification.

NRCs were responsible for finalizing their instruments and thus had to review the feedback carefully and consider the corrections and suggestions made by the verifier. The NRCs were required to review and resolve every comment the verifier made within the AM Designer, MS Word files, the NAF, and TVFs (where applicable) by accepting, modifying, or rejecting the verifier's suggestion. For the latter two options, NRCs had to give an explanation for why they changed or disagreed with the suggestion.

For issues involving adaptation implementation and trend translations, NRCs were further requested to discuss any undocumented or unclear changes with the ISC. These discussions were recorded within the final version of the NAF and/or TVF, which was then submitted for the final step of international review, layout verification.

During translation verification, some of the most common errors identified by the verifiers included:

- Translations left in the source version (English)
- Mistranslations
- Omissions/additions of text
- Inconsistent translations within or across modules/questionnaires
- Grammatical errors
- Adaptations not implemented as documented by the country.

Due to some technical issues within the AM Designer, some translations were reverted to previous translations or the English source version during the main survey. This occurred at different stages of the instrument preparation process and affected countries' translations. Some of these reversions required additional checks, occasionally after the materials were verified and considered as final. In those cases, additional checks were done to ensure that translations remained as originally implemented by the country.

Countries experienced some difficulties in the adaptation and translation of some specific terms that related to technologies within the context questionnaires and assessment materials. Examples of this includes the familiarity of students with English/target language technological terms, the adaptation or translation of specific software tools both for assessment instruments and context questionnaires (such as Limnu, Stormboard, Google Jamboard, Microsoft Whiteboard) as well as the adaptation constructs that related to the CT modules (such as commands or code blocks). Any adaptations reported in the NAF were also reviewed by the verifiers, who were asked to comment on their adequacy.

Once translation verification was completed, NRCs were asked to implement the verifier's feedback. If there were any adaptations or translations that required final approval from the ISC, these were discussed and reported in the final version of the NAF.

Layout verification

After finalizing the translations, NRCs were required to review the layout of the student instruments and adult questionnaires before submitting them for layout verification. The primary purpose of this verification step was to ensure that the materials were displayed correctly, maintained the correct structure and functionality, and, most importantly, remained comparable to the international source.

For the student instruments, layout verification took place within the AM Designer. For the adult questionnaires, countries were required to enter their final translations within OSS. In addition to this, countries submitted an additional version of the NAF with comments on layout issues that needed to be resolved with the support of the ISC.



For the layout verification of the student instruments, the preview of the national materials in the AM Designer was checked against the international version to ensure identical layout and functionality. During this review, issues reported by the country in the NAF's layout verification comments worksheet were examined, and any additional errors identified by ISC verifiers were documented. Once layout issues were compiled, they were either resolved by the verifiers when possible or forwarded to RM for resolution. A response was added to the layout verification comments once the issue was addressed. Countries were then asked to review the materials and confirm their readiness before the ISC created a USB player or made the student instruments available online.

The most common layout issues that were identified in the student instruments were:

- Missing or additional line breaks
- Text not fitting within the pre-defined spaces
- Display of HTML tags
- Issues specific to non-Latin alphabet languages such as the need for using different sizes or fonts to correctly display special characters or for better readability
- Incorrect display of left-to-right languages
- Extra spaces
- Issues with structural changes to national content in the questionnaires.

For the layout verification of the adult questionnaires, verifiers reviewed the layout of the national online adult questionnaires by comparing the OSS version with the Word files (which contained the final translations). During this review, the issues reported by the country within the NAF were examined and resolved. Any other additional errors identified by the verifiers were corrected and also documented within the NAF for countries' information. This step also included verifying the correct technical implementation of national structural adaptations in the OSS software. Countries were then asked to review the final questionnaires online and confirm their readiness.

Chapter 6:

Sampling design and implementation

Sabine Tieck

6.1 Introduction

International comparative surveys require randomly drawn samples to ensure valid inferences from the observed (sampled) data to the population characteristics under study. Random selection of study participants is a key criterion for achieving international comparability and comparability between sub-national entities. ICILS follows an international study design that meets all Technical Standards for IEA Studies (Martin et al., 1999) to ensure high sampling quality.

This chapter introduces the sampling design and implementation of the ICILS 2023 student, teacher, and school surveys, which are largely consistent with previous cycles of ICILS. Therefore, this chapter is largely based on Chapter 6 of the ICILS 2018 technical report (Tieck, 2020), which in turn was based on Chapter 6 of the ICILS 2013 technical report (Meinck, 2015). However, an important modification was made in the within-school sampling procedure for the student survey: instead of selecting students directly from a comprehensive list of all target grade students, entire classes within the target grade were sampled. In the first two cycles of ICILS, the ICILS International Study Center (ISC) was concerned that school infrastructure (in particular the number of co-located computers and the quality of internet access) may prevent more than 20 students from simultaneously completing the ICILS instruments in some schools. The change of within-school sampling method for ICILS 2023 was made possible by improvements in school-level ICT resources over time, and reduced the administrative and organizational level within schools. It is the same within-school sampling approach used in other IEA studies which over time, have transitioned from paper-based to computer-based delivery. It is beyond the scope of the ICILS design to measure the exact impact of this change in the school-level sampling approach on the data. The impact of the change would vary across countries depending on how students are allocated to classes within countries. Given the overall sampling design in ICILS, it is likely that the change in design would not affect estimates of scale means within countries (relative to previous cycles). However, it may have a very small impact on the estimates of variance on the assumption that the variation of achievement among students within classes in schools may be less than the variation in achievement of students within grade levels in schools.

The ICILS 2023 sample design is considered a “complex” design because it incorporates multi-stage sampling, stratification, and cluster sampling. This chapter describes these features in detail, along with information on target population definitions, design features, sample sizes, and achieved design efficiency. While the focus is on the international standard sampling design, specific characteristics of each national sampling plan are presented in B.

The samples of schools, students, and teachers in each ICILS country and education system were selected or verified by IEA Hamburg in collaboration with the National Research Coordinators (NRCs) of each participating country or education system. NRCs received support for their sampling activities through a series of manuals and the IEA Windows Within-School Sampling Software (WinW3S). Additionally, the sampling referee, Marc Joncas, provided guidance on sampling methodology and reviewed and adjudicated all samples for this study.

6.2 Target population definitions

In cross-country comparative surveys, it is essential to clearly define the target population(s) under study. ICILS collected data from students, their teachers, and their schools, requiring precise definitions for all three populations. These definitions enabled ICILS NRCs to accurately identify and list the



targeted schools, students (i.e., target grade classes), and teachers from which samples were drawn.

Definition: Student Target Population

ICILS defined the target population of students as follows:

The student target population in ICILS consists of all students enrolled in the grade that represents eight years of schooling, counting from the first year of ISCED level 1,^a providing the mean age at the time of testing is at least 13.5 years. Students older than 17 years are not part of the target population.

^a ISCED stands for International Standard Classification of Education (UNESCO Institute for Statistics, 2012).

For most countries, the target grade was grade 8 or its national equivalent. To ensure international comparability, the ICILS NRCs had to specify their country's legal school entry age, the name of the target grade, and an estimate of the mean age of the students in that grade. Two countries deviated from the international target grade. Norway tested their grade 9 students, which is reflected by including the tested grade in the country name for reporting, and Romania followed the international population definition but decided to survey their grade 8 students (and their teachers) at the beginning of the (next) school year. Due to this deviation, results were annotated accordingly in the international report.

Hereafter, the term “students” is used to describe “students in the ICILS target population.”

Definition: Teacher Target Population

ICILS 2023 defined the target population of teachers as follows:

Teachers are defined as school staff members who provide student instruction through the delivery of lessons to students. Teachers may work with students as a whole class in a classroom, in small groups in resource rooms, or one-to-one inside or outside of classrooms.

The teacher target population in ICILS consists of all teachers teaching regular school subjects to students of the target grade (regardless of the subject or the number of hours taught) during the ICILS testing period and have been employed at school since the beginning of the school year, i.e., within the first four weeks of the school year.^a

^a Greece included also teachers, who joined within the first two months after the beginning of the school year as this happens quite frequently.

School staff from the following categories were not regarded as part of the target teacher population (i.e., were out of scope):

- Any school staff that attend to the needs of target grade students but do not teach any lessons (e.g., psychological counselors, chaplains);
- Assistant teachers and parent-helpers;
- Non-staff teachers who teach (non-compulsory) subjects that are not part in the curriculum (e.g., cases where religion is not a regular subject and taught by external persons);
- Teachers who have joined a school in the fifth week (or later) of the school year; and
- Teachers that are on a long-term leave during the testing period (e.g., maternity or sabbatical leave).

Hereafter the term “teachers” is used to describe “teachers in the ICILS target population.”

**Definition: Schools**

In ICILS 2023 schools were defined as follows:

A school is one whole unit with a defined number of teachers and students, which can include different programs or tracks. The definition of 'school' should be based on the environment that is shared by students, which is usually a shared faculty, set of buildings, social space and also often includes a shared administration and charter.

Schools eligible for ICILS are those at which target grade students are enrolled.

In order to ensure international comparability, the definition of "school" should be equivalent in all participating countries. In most cases, identifying schools for sampling purposes in ICILS was straightforward. However, there were some cases where identification of schools for sampling purposes was more difficult. National centers were provided with the following examples in order to help them identify sampling units:

- A multi-shift school that consists of two units that share the same building should be regarded as two different schools as long as these shifts operate as different schools (with different staff) and share only the same building (and administrative unit). Both units should be listed as schools in the School Sampling Frame;
- A multi-shift school in which students are assigned to different shifts but otherwise functions as a single entity (in terms of staffing, enrollment, and administration) should be regarded as one school and should be listed only once in the frame;
- The parts of a campus school (larger administrative schools with sub-units scattered around cities or regions) should be regarded as different schools for sampling. However, if a part of a campus school is selected for ICILS, the principal and the ICT coordinator of the combined school will be asked to fill out the school questionnaire and will need to be instructed to respond only with respect to the sampled sub-unit;
- Schools that consist of two administrative units but have shared staff shared buildings, and offer some opportunities for the students to change from one school to the other should be regarded as one combined school;
- The parts of a school with two or more different study programs (e.g., a vocational program and an academic program) that have different teaching staff, are conducted in different buildings and offer no or very little opportunity for students to change from one study program to the other, should be regarded as two or more different schools for sampling. Those study programs should be listed in the School Sampling Frame; and
- Schools that provide distance learning to students of the target population on a regular basis are considered to be schools according to the above definition.

Schools that are not intended to be included in ICILS are considered out of scope. Examples include:

- Schools without students in the target grade; and
- Schools in which all students in the target grade do not fit the age definition for students in ICILS, e.g., evening schools for adults.

6.3 Coverage and exclusions

The ISC encouraged participating countries to include all schools, students, and teachers defined in the international target populations in the study in order to ensure full coverage of these target populations, see figure 6.1. For ICILS 2023, in all but one country, the international desired target population equals the national desired target population, in other words, all parts of the population were covered, see below for details.

Figure 6.1: Target Population - Internationals desired to National defined

International Desired Target Population		
National Desired Target Population		Not covered
National Defined Target Population	School-level exclusions	
	Within-sample exclusion	

However, it was deemed appropriate to exclude some schools, students, and teachers from the respective target population for practical reasons, such as difficult test conditions or prohibitive survey costs, e.g., in remote regions. Some students and teachers were not surveyed due to the exclusion of their entire school from the sampling frame (school-level exclusions). Additionally, some classes or students (but not teachers) were excluded from participating schools (within-sample exclusions). The amount of the overall exclusion rate for the student population is limited to five percent after rounding to the nearest integer. Countries exceeding this limit are annotated in the respective tables of the international report. Fifteen countries had an overall exclusion rate between five percent and ten percent of their respective national desired target population.

As in other large-scale assessments, it should be emphasized that the ICILS 2023 sample data represents the nationally defined target populations only, i.e., without the excluded members of the nationally desired target population.

Population coverage

If a significant portion of the international desired target population must be excluded, it is reported as reduced coverage. This typically occurs when certain regions cannot be assessed (e.g., war zones) or specific types of schools are not included (e.g., minority language schools). In ICILS 2023, all but one country achieved full coverage of their nationally desired target populations. Bosnia and Herzegovina has a reduced coverage of 60.8 percent because students from the Republika Srpska did not participate. The sample for Bosnia and Herzegovina includes students from the Federation of Bosnia and Herzegovina and the Brčko District.

Exclusions of schools

In ICILS 2023 in most countries, very small schools and schools exclusively dedicated to students with special needs students were excluded. Frequently, schools providing lessons in other language of instructions or following a curriculum that differed from the mainstream curriculum were also not part of the nationally defined target population.

Table 6.1 provides an overview of the typical school-level exclusions applied in ICILS 2023, along with their respective percentages relative to the desired national school target population in each participating country. Schools identified for exclusion were removed from the school sampling frame before the sampling process. The *percentage of excluded schools* (ER_{sch}) were calculated as:

$$ER_{sch} = \frac{E_{sch}}{TP_{sch}}, \quad (6.1)$$



where E_{sch} denotes the number of schools excluded prior to sample selection, and TP_{sch} is the total number of schools belonging to the national desired target population. The respective figures were provided by the NRCs.

Because school-level data (collected via the principal and ICT coordinator questionnaires) in the ICILS 2023 survey were only used to complement the reporting of student- and teacher-level data, no specific thresholds were determined for exclusions at the school level. However, the percentages of students and teachers excluded due to the removal of entire schools were considered when determining the overall proportions of students (see below).

School exclusions varied significantly across countries, which should be considered when interpreting results based on school-level data. Additionally, since school exclusions typically affect smaller schools, the percentage of excluded schools tends to be higher than the corresponding percentages of excluded students or teachers.



Table 6.1: Percentages of schools excluded from the ICILS target populations

Country	School-level exclusion (% of all schools)					Within-sample* exclusion (%)	Overall school exclusion (%)
	Very small schools	Special need schools	Language of instruction	Other	Total		
Austria	2.3	13.5	-	0.1	15.9	1.1	16.9
Azerbaijan	10.7	0.6	0.7	-	12.0	-	12.0
Belgium (Flemish)	6.4	8.0	-	-	14.3	-	14.3
Bosnia and Herzegovina	-	3.4	-	3.4	6.9	-	6.9
Chile	4.1	0.1	-	0.1	4.2	-	4.2
Chinese Taipei	0.8	2.7	-	0.6	4.2	-	4.2
Croatia	-	1.9	-	1.0	2.9	-	2.9
Cyprus	1.0	1.0	2.9	1.0	5.7	-	5.7
Czech Republic	0.5	11.6	0.3	-	12.5	-	12.5
Denmark	1.6	16.8	0.8	-	19.2	-	19.2
Finland	-	5.6	-	1.0	6.7	-	6.7
France	0.1	0.8	-	10.8	11.6	-	11.6
Germany	1.9	9.1	-	0.0	11.1	-	11.1
Greece	3.4	3.5	-	-	6.9	-	6.9
Hungary	-	6.4	0.6	-	7.0	-	7.0
Italy	0.3	0.0	1.3	-	1.6	-	1.6
Kazakhstan	12.0	1.3	0.5	0.3	14.1	5.6	19.8
Korea, Republic of	3.2	4.6	-	5.3	13.1	-	13.1
Kosovo	-	0.6	7.4	-	8.0	-	8.0
Latvia	-	7.4	1.1	1.6	10.0	0.6	10.6
Luxembourg	-	-	-	-	0.0	-	0.0
Malta	4.2	4.2	4.2	-	12.5	-	12.5
Netherlands	0.4	20.6	0.8	2.4	24.1	-	24.1
Norway (Grade 9)	10.2	2.6	2.3	0.4	15.5	-	15.5
Oman	3.7	0.3	-	3.3	7.3	3.8	11.1
Portugal	0.2	1.8	-	6.0	8.0	-	8.0
Romania	3.4	4.0	-	2.5	9.9	4.4	14.3
Serbia	3.3	4.1	5.4	-	12.8	-	12.8
Slovak Republic	2.4	0.8	0.4	-	3.7	-	3.7
Slovenia	-	6.9	0.6	1.2	8.8	-	8.8
Spain	0.5	1.9	-	1.8	4.2	-	4.2
Sweden	2.7	13.7	-	1.3	17.7	-	17.7
United States	-	-	-	-	0.0	-	0.0
Uruguay	-	0.2	-	7.7	7.8	-	7.8
Benchmarking participants							
North Rhine- Westphalia, Germany	0.4	11.1	-		11.5	-	11.5

*Within-sample exclusion consists here of sampled schools with excluded students only.

Note: (a) Figures are rounded. (b) School exclusion percentage is not related to the percentage of excluded students.

Please note that the school-level exclusion rate for the student population is calculated as the proportion of students in the excluded schools. In addition, the proportion of students in very small schools must not exceed two percent of the national desired target population. Therefore, the definition of a very small school depends on the country's data.

In rare cases, schools that should have been excluded were not marked as such in the school frame and were therefore not removed prior to school sampling and were not included in the school-level exclusion rate. If such schools were sampled, they contributed to the within-sample exclusion rate.

Student-level exclusions

Students can be excluded by excluding schools (school-level exclusion) and they can be excluded within participating schools (within-sample exclusion). However, within-sample exclusion should be minimized and is permitted according to the following international within-school exclusion rules:

- Students with an intellectual disability – These are students who are considered, in the professional opinion of the school principal or by other qualified staff members, to have an intellectual disability or who have been tested psychologically and reported as having an intellectual disability. This includes students who are emotionally or mentally unable to follow even the general instructions of the test. Students should not be excluded solely because of poor academic performance or normal disciplinary problems;
- Students with a functional disability – These are students who have a permanent physical disability that prevents them from participating in the ICILS testing. Functionally disabled students who are able to respond should be included in the testing, and
- Non-native language speakers – These are students who are unable to read or speak the language(s) of the test and would be unable to overcome the language barrier of the test. Typically, a student who has received less than one year of instruction in the language(s) of the test should be excluded, but this definition may need to be adapted in different countries.

In the case where classes within sampled schools consist entirely of students who belong to one of the exclusion categories, these classes can be excluded prior to class sampling.

Finally, the students' overall exclusion rate is computed as the sum of the students' school-level exclusion rate and the weighted within-sample exclusion rate. Table 6.1 and B of this report provide details about the exclusion types for each country. Regarding the overall rate of excluded students, each country was required to keep the overall rate (due to school-level and within-sample exclusions) below five percent (after rounding) of the respective desired student target population.

Table 6.2 provides the percentages of coverage and exclusions for the ICILS 2023 student population for each country participating in ICILS 2023. In 15 education systems participating in ICILS 2023, the overall student exclusion rate was above five percent but less than ten percent, which resulted in respective annotations in the ICILS 2023 international report (Fraillon, 2025). Why such a high number of participating countries exceeded the allowed is not part of this chapter. One country, Bosnia and Herzegovina, covered only 61 percent of their entire student population. Of this remaining student population, another five percent of students were excluded.



Table 6.2: Percentages of students covered and excluded from the ICILS student target population

Country	Coverage: int. target population (%)	Exclusion from student target population (%)			
		School-level exclusions	Within-sample exclusions* Language issues	Total	Overall exclusion (%)
Austria	100	2.1	2.8	5.4	7.6
Azerbaijan	100	1.9	0.0	0.3	2.2
Belgium (Flemish)	100	1.4	1.0	1.3	2.7
Bosnia and Herzegovina	60.8	3.9	0.0	1.2	5.2
Chile	100	0.3	0.0	0.9	1.2
Chinese Taipei	100	0.3	0.1	1.8	2.1
Croatia	100	0.6	0.9	5.2	5.8
Cyprus	100	0.9	0.9	2.0	2.9
Czech Republic	100	3.0	2.7	3.1	6.2
Denmark	100	3.7	0.1	3.2	6.8
Finland	100	1.3	1.3	2.8	4.2
France	100	3.0	0.7	1.8	4.8
Germany	100	1.6	1.8	2.3	3.9
Greece	100	0.7	1.2	2.3	3.0
Hungary	100	2.4	0.6	2.1	4.5
Italy	100	0.8	0.0	3.8	4.7
Kazakhstan	100	2.2	3.2	4.0	6.1
Korea, Republic of	100	1.7	0.4	1.5	3.2
Kosovo	100	4.9	1.0	1.6	6.5
Latvia	100	5.7	2.3	4.0	9.7
Luxembourg	100	3.6	0.8	1.4	5.0
Malta	100	1.4	0.1	2.8	4.2
Netherlands	100	4.5	0.1	0.9	5.4
Norway (Grade 9)	100	1.9	0.0	4.4	6.3
Oman	100	1.1	0.1	0.8	1.9
Portugal	100	6.2	1.0	2.1	8.2
Romania	100	3.8	1.2	3.4	7.2
Serbia	100	3.8	2.4	2.8	6.6
Slovak Republic	100	0.6	1.4	2.8	3.3
Slovenia	100	2.8	1.2	3.5	6.3
Spain	100	1.3	1.9	4.5	5.8
Sweden	100	1.6	2.3	7.1	8.7
United States	100	0.0	1.0	3.5	3.5
Uruguay	100	0.9	0.0	1.2	2.2
Benchmarking participants					
North Rhine-Westphalia, Germany	100	1.7	1.6	2.0	3.7

*Within-sample exclusion rates are estimates based on sample data. Some participating countries reported an increasing number of refugees and that this led to an increase of the within-sample exclusion. Therefore it deemed necessary to calculate this estimate and to provide the within sample exclusion rate due to language issues.



The student's school-level exclusions consisted of those students belonging to schools which were excluded prior to the school sampling. The students' school-level exclusion rate ER_1 was calculated as:

$$ER_1 = \frac{E_1}{TP}, \quad (6.2)$$

where E_1 denotes the number of target grade students in excluded schools, and TP is the total number of students belonging to the national desired target population. The respective figures were provided by the NRCs.

The students within-sample exclusions were based on information collected from the sample (i.e., after the school sampling step). The within-sample exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test. Complete classes with to-be-excluded students only could be excluded prior to the within-school sampling; single students could be excluded after the within-school sampling (class sampling) was performed.¹² The percentage of student within-sample exclusions was calculated using the estimated number of students excluded within sampled schools and the total number of students belonging to the national desired target population.

The students' within-sample exclusions ER_2 were computed as:

$$ER_2 = \frac{W^{E_2}}{W^P + W^{E_2}} \times (1 - ER_1), \quad (6.3)$$

where W^{E_2} is the sum of weights of excluded students identified in the sample and W^P is the sum of weights of participating students. Therefore, $W^P + W^{E_2}$ denotes the (estimated) total number of students belonging to the nationally desired target population. To make inferences not only to the sample, but to the student population the students' school-level exclusion rate is taken into account by multiplying the fraction by $1 - ER_1$.

The overall exclusion rate of students ER_{tot} is the sum of the students' school-level exclusion rate and the weighted within-sample exclusion rate:

$$ER_{tot} = ER_1 + ER_2. \quad (6.4)$$

Teacher-level exclusions

Teachers working in excluded schools were not part of the nationally defined target population. Within participating schools, all teachers who met the target population definition were eligible for participation in the survey.

Each country was asked to provide information on the total number of teachers teaching in the target grade, as well as the proportion of these teachers in excluded schools. For 19 countries, no statistics on the number of eligible ICILS 2023 teachers were available, making it impossible to compute exclusion rates. Teacher data were included in the ICILS 2023 international report (Fraillon, 2025) only if the student school-level exclusion rate exceeded five percent after rounding. In such cases, it is expected that more than five percent of teachers were also excluded.

¹² In some cases, these students were grouped in classes, which were then excluded as a group, or the sample schools were found to have only enrolled students within the exclusion categories, which resulted in the corresponding school(s) being excluded ex-post.



6.4 School sampling design

For all ICILS 2023 countries, a stratified two-stage probability cluster sampling design was used to conduct school sample selection. In the first stage, schools were systematically selected with probabilities proportional to their size (PPS), where 'school size' was measured by the total number of enrolled target grade students. In the second stage, within participating schools, one or more classes from the target grade were selected using a systematic simple random sampling approach. All students in the selected classes were invited to participate.

The following subsections provide further details on the school sample design for ICILS 2023.

School sampling frame

In order to prepare the selection of school samples, national centers provided a comprehensive list of schools, including the number of students enrolled in the target grade. This list is referred to as the school sampling frame. To ensure that each ICILS 2023 school sampling frame provided complete coverage of the desired target population, the sampling team carefully checked and verified the plausibility of the information by comparing it with official statistics.

The sampling team required the following information for each school in the sampling frame:

- Random school IDs,¹³ which are unique and used solely for the purpose of the sample selection—for instance, a sequential number;
- School's measure of size (MOS), which was usually the number of students enrolled in the target grade or an adjacent grade;
- Values for each of the intended stratification variables; and
- A school exclusion variable, which indicates the schools to be excluded.

Stratification of schools

Stratification is part of many sampling designs and entails grouping sampling frame units by common characteristics. Examples of such groups of units (schools in the case of ICILS 2023) would be performance level, administration type, geographic region, urbanization level, or source of funding.

ICILS 2023 applied two different methods of stratification. Implicit stratification means that the sampling frame has been sorted, prior to sampling, by implicit stratification variables, thus providing a simple and straightforward method to achieve a fairly proportional sample allocation across all strata. With explicit stratification, independent samples of schools are selected from each explicit stratum. The sample sizes for each explicit stratum are assigned before the selection process in order to achieve the desired sample precision overall and, where required, also for subpopulations.

In summary, IEA studies use stratification for the following reasons:

- To improve the efficiency of the sample design, i.e., reducing standard errors, thereby making survey estimates more reliable. For this, national centers identify stratification variables expected to be closely associated with students' learning outcomes. This can be achieved using either implicit or explicit stratification.
- To apply optimized sample designs, such as disproportionate sample allocations to specific groups of schools (e.g., states or provinces). This is particularly important when subgroup estimates are required by the national analysis plan. Explicit stratification is used for this purpose.
- To ensure a proportional representation of specific groups of interest (domains) of the target population in the sample. This is achieved by applying implicit stratification.

¹³ Due to GDPR, national school IDs cannot be used anymore within Europe and shouldn't be used in general.



The latter design feature was used if the country required estimates with higher precision levels for specific subgroups of interest in the target population. For example, a country may wish to compare public and private schools, but only 10 percent of the students in that country attend schools in the latter category. In such a case, a proportional sample allocation would result in having too few private schools in the sample to provide reliable estimates for this particular subpopulation, and even larger differences between the two different school types might not appear as statistically significant.

To allow comparisons with sufficient statistical power without introducing any sample bias, an appropriately larger sample of private schools could be selected while keeping the original proportional sample allocation for public schools the same. In the example above (with a distribution of 10 percent of students in private and 90 percent in public schools), a proportional sample allocation for a selection of 150 schools would lead to a sample of 15 private and 135 public schools. Through oversampling, it would be possible to increase the number of private schools to a sufficiently large number (for example, by selecting 50 private and 135 public schools).

Each country applied different stratification schemes following discussions with IEA sampling experts. Table 6.3 provides details about the stratification variables used.



Table 6.3: Stratification schemes of participating countries

Country	Explicit stratification variables (Number of variable characteristics)	Number of explicit strata	Implicit stratification variables (Number of variable characteristics)
Austria	School Type (3), Individual Student Funding (2), Urbanization (2)	7	Federal States (9)
Azerbaijan	Language (2), Urbanization (4)	5	Urbanization for Azerbaijani and Russian Schools (4)
Belgium (Flemish)	School Type (4)	4	School Administration Type (2)
Bosnia and Herzegovina	Canton (10)	10	None
Chile	School Administration Type (3)	3	National Assessment Performance Group for Mathematics (4)
Chinese Taipei	Urbanization (5), Region (2)	8	School Performance (5)
Croatia	Region (6)	6	None
Cyprus	District (4), School Administration Type (2), Curriculum (2)	5	Urbanization (2)
Czech Republic	School Type (2), District (14)	15	District for School Type Multi-year Gymnasium (14)
Denmark	None	1	Achievement Score (Grade Point Average) (5)
Finland	School Language (2), Regional State Administrative Agency (4), Urbanization (2)	8	Regional State Administrative Agency for Swedish Speaking Schools (6)
France	School Administration Type (3)	3	None
Germany	German State (2), School Type (3)	5	German State (16), Socioeconomic Status (Migrant Proportion) (50)
Greece	School Administration Type (2), Administrative Regions (14)	15	Urbanization (5)
Hungary	Urbanization (4), National Assessment Mathematics Score (4)	10	Region (7)
Italy	Region (6)	6	School Administration Type (2), National Test Results (3)
Kazakhstan	Region (17)	17	Urbanization (2), School Language (3)
Korea, Republic of	Urbanization (3), School Type by Gender (3)	9	City (17)
Kosovo	Region (7)	7	Urbanization (2)
Latvia	Urbanization (3)	3	School Type (3)
Luxembourg	School Administration Type (3)	3	None
Malta	Sector (3)	3	None
Netherlands	Educational Tracks (3)	3	None
Norway (Grade 9)	Region (2)	2	Language (2), National Examination School Score (4)
Oman	Governorate (11), School Administration Type (2)	13	School Type (3)
Portugal	Grade (2), School Administration Type (2), Region (6)	9	School Administration Type (2), Region (5)
Romania	Region (5) And Urbanization (Rural, Urban)	8	Urbanization (2)
Serbia	Region (3), Urbanization (2), School Hierarchy (2)	7	National Assessment Performance Group for Mathematics (4)
Slovak Republic	Language (3), National Math Assessment (3)	5	School Type (3), National Math Assessment (4)
Slovenia	None	1	None
Spain	Region (12)	12	School Administration Type (2), Autonomous Communities (8)

Table 6.3: Stratification schemes of participating countries (cont'd)

Country	Explicit stratification variables (Number of variable characteristics)	Number of explicit strata	Implicit stratification variables (Number of variable characteristics)
Sweden	Average Grades (4), School Administration Type (2)	7	Administration Type for Schools with Missing Average Grade (2)
United States	School Administration Type (2), Region (3), Poverty Level (2)	8	Urbanization (4), Percentage of Non-White Students (2), Federal Interagency Panel on School Funding (51)
Uruguay	School Administration Type (2), Learning System (2), Region (2)	6	None
Benchmarking participants			
North-Rhine Westphalia, Germany	School Type (3)	3	Socioeconomic Status (3)

Note: Please note that the number of explicit strata in the third column is not automatically the product of the levels of the stratification variables in column 2. For example, a country with two explicit stratification variables (region - five level, and urbanization - two level) ends up with eight strata as one region has only school located in urban environment, while another region has only schools in rural environment. For country details on the applied stratification, see [Appendix B](#).

School sample selection

To select school samples for the ICILS 2023 main survey, the sampling team used stratified PPS systematic sampling. This method is standard in most large-scale educational surveys, including most IEA surveys. Under ideal conditions—that is, in the absence of non-response and disproportional sampling of subpopulations—this method would lead to “self-weighted” samples where all units sampled in the last stage of sampling have a similar probability of selection. In reality, however, not a single ICILS sample was self-weighting. First, note that a sampling design can only be self-weighting for one target population, and ICILS is aimed at self-weighted samples of students. In turn, and by design, the samples of schools and teachers cannot be self-weighting. Second, for most countries, the implemented design actually guaranteed that the samples were not self-weighting because explicit stratification was used.¹⁴

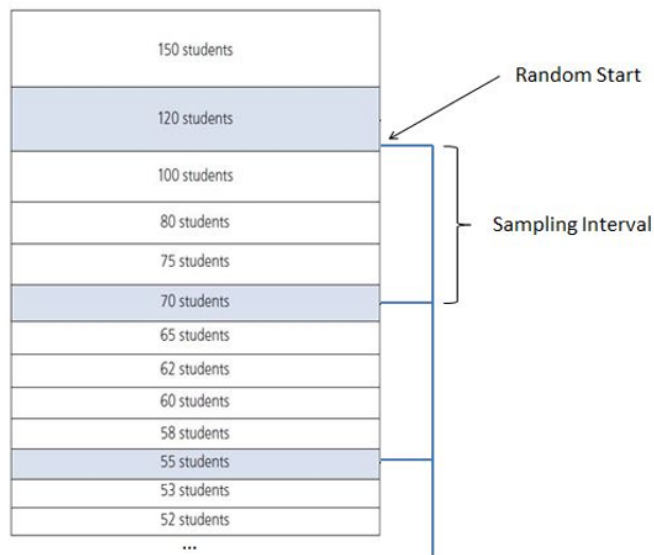
General school sample selection involved the following steps:

1. Splitting the school sampling frame containing all eligible schools into separate frames according to the defined, explicit strata unless no explicit stratification was used. In the latter, all schools from the school sampling frame were kept in one explicit stratum. All of the following steps were done independently within each explicit stratum.
2. Sorting the schools by implicit strata and by MOS (alternately sorted in increasing and decreasing order).
3. Calculating a sampling interval (SI) by dividing the total MOS by the number of schools to be sampled from that explicit stratum.
4. Identify the first sampled school by multiplying the sampling interval with a random number (=random start).
5. For the n^{th} school to be selected, calculate the value for selection by adding $(n - 1) * SI$ to the random start. Whenever the cumulated MOS is equal to or above the value for selection, the corresponding school was included in the sample.

¹⁴ When using explicit stratification, sampling allocation can hardly ever be exactly proportional for different explicit strata. In addition, for really small explicit strata, a minimum number of schools needs to be sampled, which then leads to the oversampling of such schools.

Figure 6.2 visualizes the process of systematic PPS sampling within an explicit stratum. In this diagram, the schools in the sampling frame are sorted descending by MOS, and the height of the cells reflects the number of target grade students in each school. A random start determines the second school in the list for selection, and a constant sampling interval determines the next sampled schools. Cells with sampled schools are shown as shaded. Joncas and Foy (2012) provide a more comprehensive description of the sampling process using an illustrative example.

Figure 6.2: Visualization of PPS systematic sampling.



In the following specific cases, the sampling design deviated from this general procedure:

- Small schools were selected with equal selection probabilities to avoid large variations of sampling weights due to expected changing size measures. A school was regarded as “small” if the number of target grade students was less than the minimum cluster size¹⁵ in the respective explicit stratum.
- Very large schools (i.e., schools with more students than the value of the sampling interval) were temporarily removed from the explicit stratum and selected with certainty (i.e., all schools in this category were included in the sample).

All countries conducted a field trial with a small sample of schools one year before the ICILS main survey. Where possible it is preferable that any given school is not selected to participate in both the field trial and the main survey data collection. This is because selection in both the field trial and main survey may reduce the likelihood of a school agreeing to participate and also because there may be some information sharing within a school about the contents of the instruments that could influence the data collected in the main survey. In ICILS, this was prevented by selecting the sample of schools to participate in the field trial and the main survey simultaneously (as part of a single larger sampling procedure) in each country. For example, if a country was planning to sample 25 schools in the field trial and 150 schools in the main survey, then a single combined sample of 175 schools was selected first. Then, from these schools, the main survey sample of 150 schools was subsampled, leaving the remaining (25) schools for the field trial sample.

ICILS 2023 was conducted in the same year as the IEA Trends in International Mathematics and Science

¹⁵ The minimum cluster size is calculated as the average class size of all schools with more than one class.

Study (TIMSS) 2023. A lot of countries requested that schools be selected for one of these studies only to reduce the administrative burden for the schools. The ICILS 2023 sampling team collaborated closely with the sampling colleagues, implementing the sampling for TIMSS 2023 to prevent school sample overlap whenever possible. The procedures used to prevent school overlap ensured randomness of selection, known (correct) school selection probabilities, and consequently unbiased samples for all of these studies. Where it was not possible to achieve distinct school samples for both studies, different classes were selected randomly for both surveys where possible.

Once schools had been selected from the sampling frame, up to two (non-sampled) replacement schools were assigned for each originally sampled school. In order to reduce the risk of non-response bias due to replacement, the replacement schools were usually assigned as follows: The school that appeared directly after a selected school in the sorted sampling frame was assigned as its first replacement, while the preceding school of the sampled school was used as its second replacement. This ensured that replacement schools shared similar characteristics with the corresponding sampled schools to which they belonged, were of similar size, and belonged to the same stratum. Please note that—as in all IEA studies—the use of replacement schools in ICILS 2023 was limited; see ICILS 2023 standards for sampling participation in 7.

6.5 Within-school sampling design

Within-school sampling constituted the second stage of the ICILS sampling process. The NRCs or their appointed data managers carried out the selection of students and teachers. The use of the WinW3S software, developed by IEA Hamburg, ensured the random selection of classes (and therefore students) and teachers within the sampled schools. Replacement of non-responding individuals was not permitted in ICILS 2023.

Class sampling

WinW3S employed systematic stratified sampling with equal selection probabilities to select one or more classes from comprehensive lists of classes with target grade students provided by the participating schools. Explicit or implicit stratification, e.g., by track, could be applied. This was applied in close collaboration with the IEA sampling team. In cases where the school was also selected for TIMSS 2023 and has two or more target grade classes, a different class was randomly selected for ICILS 2023.

Teacher sampling

WinW3S also employed systematic stratified sampling with equal selection probabilities to select teachers from comprehensive lists of in-scope teachers provided by the participating schools. The procedure also ensured a sample allocation among subgroups that was near to proportional. To increase sample precision, teacher lists were implicitly stratified by sorting them according to gender, main subject domain, and birth year.

6.6 Sample size requirements

The ISC, following standard practice in other IEA studies, set high standards for sampling precision and aimed to achieve reasonably small standard errors for survey estimates. The student sample was designed to ensure a specified level of precision for population estimates, defined by confidence intervals of ± 0.1 standard deviation for means and $\pm 5\%$ for percentages. With respect to the main outcome variable in ICILS 2023—that is, the students' computer and information literacy (CIL) score scale established in ICILS 2013 with a mean of 500 score points and a standard deviation of 100 for equally weighted national samples from the first cycle—this requirement translated into standard errors that needed to be below five score points. IEA was responsible for determining sample sizes that were expected to meet these requirements for each participating country. With the exception of three participating countries (United States, Romania, and Azerbaijan), which failed to meet the IEA sample participation standards, all participating countries and education systems achieved this requirement (see Chapter 5 in Fraillon, 2025). The required precision levels of percentages were also met for the vast majority of population estimates presented in the ICILS international report.

There were also other considerations that needed to be taken into account when determining the required number of sampled students and teachers:

- Some types of analysis, like multilevel modeling, require a minimum number of valid cases at each sampling stage (see, for example, Mang et al., 2021);
- For the purpose of building scales and sub-scales, a minimum number of valid entries per response item is required; and
- Reporting on subgroups (e.g., age or gender groups) requires a minimum sample size for each of the subgroups of interest.

All these considerations were taken into account during the process of defining minimum sample sizes for schools, students, and teachers.

School sample sizes

The minimum sample size for the ICILS main survey was 150 schools for each country. In some countries, it was necessary to select more schools than the minimum sample size due to one or more of the following reasons:

- Previous student surveys had shown a relatively large variation in student achievement between schools in a country. In these cases, it was assumed that the IEA standards for sampling precision could only be met by increasing the school sample size.
- The number of “small schools” in the target grade was relatively large, and it was not possible to reach the student sample size requirements by selecting only 150 schools (see next section below).
- The country requested oversampling of particular subgroups of schools to accommodate national research interests.

Student sample sizes

In most sampled schools, one class was randomly selected from the full list of classes containing target grade students. Each country was required to achieve a student sample size of approximately 3,000 tested students, which corresponds to around 3,350 to 3,400 sampled students. However, due to factors such as non-response, changes in school size, school closures, or other circumstances, some countries did not reach this target. The ICILS 2023 sampling team did not consider this outcome problematic as long as the country met the overall participation rate requirements.

Teacher sample sizes

Typically, 15 teachers of the target grade were randomly selected in each sampled school. In schools with 20 or fewer teachers of the target grade, all teachers were selected. Countries with less than 150 schools were asked to increase the number of teachers to sample in each school.

In summary, the minimum sample size requirements per country for ICILS 2023 were as follows:

- Schools: 150 (or all schools if fewer than 150 exist in the country)
- Classes: at least one class per school
- Students: a total of 3,350 to 3,400 sampled students, aiming for at least 3,000 participating students
- Teachers: 15 per school (or all, if 20 or fewer)

Table 6.4 lists the intended and achieved school sample sizes, the achieved student sample sizes, and the achieved teacher sample sizes for each participating country. Note that schools may have been



treated as participating in the student survey but not in the teacher survey and vice versa due to specific minimum within-school response rate requirements. This explains the differences in the number of participating schools in the student and teacher survey across ICILS 2023 countries.

Table 6.4: School, student, and teacher sample sizes - expected versus realized

Country	Originally sampled schools ¹	Expected number of listed students ²	Student survey		Teacher survey	
			Participating schools	Participating students	Participating schools	Participating teachers
Austria	160	3547	154	3448	133	1630
Azerbaijan	164	3565	164	3634	147	2202
Belgium (Flemish)	154	3468	136	3365	124	1415
Bosnia and Herzegovina*	107	2111	105	1877	38	451
Chile	168	5344	129	3216	60	528
Chinese Taipei	169	5345	169	5112	167	2233
Croatia	152	3366	148	2911	147	1903
Cyprus	99	3567	95	3182	89	1036
Czech Republic	210	8586	210	8169	210	3117
Denmark	150	3811	141	3038	125	979
Finland	158	4178	158	4249	155	2099
France	152	3871	150	3694	119	1504
Germany	230	5706	222	5065	180	2302
Greece	186	4164	179	3576	172	2321
Hungary	158	4083	155	3491	142	1596
Italy	152	3587	152	3376	151	2161
Kazakhstan	178	4008	176	4852	176	2567
Korea, Republic of	152	4060	152	3723	150	2054
Kosovo	154	3640	153	3345	110	1197
Latvia	151	3358	143	2705	133	1674
Luxembourg**	47	7229	46	4703	20	309
Malta	42	2285	42	3115	31	405
Netherlands	155	4224	46	1288	30	300
Norway	160	3953	153	4436	140	1366
Oman	222	6513	221	6437	218	2661
Portugal	164	5759	164	3650	164	2089
Romania	156	4387	136	3270	130	1678
Serbia	156	3452	154	3125	155	2224
Slovak Republic	166	3565	166	3034	164	2330
Slovenia	175	3815	168	3318	165	2329
Spain	510	13428	508	11799	500	6247



Table 6.4: School, student, and teacher sample sizes - expected versus realized (cont'd)

Country	Originally sampled schools ¹	Expected number of listed students ²	Student survey		Teacher survey	
			Participating schools	Participating students	Participating schools	Participating teachers
Sweden	154	3922	147	3401	143	1566
United States***	186	4650**	118	2352	95	898
Uruguay	154	4317	144	2933	129	1464
Benchmarking participants						
North Rhine-Westphalia, Germany	113	2935	111	2726	106	1459

¹Please note that during the main survey data collection some sampled schools were discovered to be ineligible or excluded.

²The number of expected listed students is based on the information on the school frame provided, and should be large enough to cover ineligible or excluded schools as well as expected non-response on school- or student level.

* Note reduced coverage of 61%.

** One school was split into two virtual schools for better representation of offered curricula.

*** Figure based on an average class size of 25 students.

References

- Fraillon, J. (Ed.). (2025). *An international perspective on digital literacy: Results from ICILS 2023*. Springer. <https://doi.org/10.1007/978-3-031-87722-3>
- Joncas, M., & Foy, P. (2012). Sample design and implementation. In M. Martin & I. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/methods/pdf/Sampling_Schools.pdf
- Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-scale Assessments in Education*, 9(6). <https://doi.org/10.1186/s40536-021-00099-0>
- Martin, M. O., Rust, K., & Adams, R. J. (1999). *Technical standards for IEA studies*. International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/iea-reference/technical-standards-iea-studies>
- Meinck, S. (2015). Sampling design and implementation. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *IEA International Computer and Information Literacy Study 2013 technical report* (pp. 67–86). International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2013-technical-report>
- Tieck, S. (2020). Sampling design and implementation. In J. Fraillon, J. Ainley, W. Schulz, T. Friedman, & D. Duckworth (Eds.), *IEA International Computer and Information Literacy Study 2018 technical report* (pp. 59–78). International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>
- UNESCO Institute for Statistics. (2012). *International Standard Classification of Education: ISCED 2011*. UNESCO-UIS. <https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf>

Chapter 7:

Sampling weights and participation rates

Sabine Tieck

7.1 Introduction

One major objective of ICILS 2023 was to make use of data collected from samples of students, teachers, and schools within each country (or education system) to make accurate, precise, and internationally comparable estimates of population characteristics. In order to achieve this objective, it is important to consider the particular feature of the study's complex sample design (see [Chapter 6](#) for details). Data from complex samples may provide biased population estimates if analysts disregard the particular features of complex samples and treat data as if they were from a simple random sample.

One particularly important feature of a complex sample is that sampling units do not have equal selection probabilities. In ICILS 2023, this characteristic applied to the sampled students, teachers, and schools. Furthermore, non-participation can potentially lead to bias estimates, and differential patterns of non-response increase this risk. To account for these complexities, sampling weights and non-response adjustments were computed within each participating country, leading to an estimation (or "final") weight for each sampled unit.¹⁶ All findings presented in ICILS 2023 reports are based on weighted data.

This chapter is largely based on Chapter 7 of the ICILS 2018 technical report (Tieck, [2020](#)), which in turn was based on Chapter 7 of the ICILS 2013 technical report (Meinck & Cortes, [2015](#)). Descriptions of how the several sets of weights and non-response adjustments were computed follow. The conditions under which students, teachers, and schools were judged to be "participants" are described within the relevant sections describing the weights for each of these groups. Subsequent sections in this chapter describe the computation of participation rates at each sampling stage, and the minimum participation requirements. The ICILS 2023 ISC regarded response rates as an important indicator of data quality and the achieved quality of sample implementation for each country is presented. Please note that [Chapter 13](#) covers the use of the jackknife replication method (needed for variance estimation).

7.2 Types of sampling weights

The ICILS 2023 *final weights* are the product of several weight components. Generally, it is possible to discriminate between two different types of weight components:

- *Base (or design) weights*: these reflect selection probabilities of sampled units. They are computed separately for each sampling stage and therefore account for multiple-stage sampling designs (see [Chapter 6](#) for details). The base weight of a sampled unit is the inverse of the product of the selection probabilities at every stage.
- *Non-response adjustments*: these aim to compensate the potential for bias due to nonparticipation of sampled units. As with base weights, they are computed separately for each sampling stage. The main function of this weight component is that the (base) weight of the non-respondents within a specific adjustment cell is redistributed among the responding units in that cell. Such an "adjustment cell" contains sampling units that share specific features. For example, all private schools in a given region could comprise a stratum of schools, from which a sample of

16 For further reading on the topic, we recommend Franklin and Walker ([2003](#)), Groves et al. ([2004](#)), Meinck ([2015](#)), and Rust ([2014](#)).



schools was selected. If some of the sampled schools did not participate, then the remaining (participating) schools in this stratum would carry the (base) weight of the non-participating schools. This approach allows us to exploit the (usually) little information we have available about respondents and non-respondents, and to assume that school non-participation is associated with the different strata (see also Lohr, 1999). The approach also assumes a non-informative response model, implying that non-response occurs completely at random within the adjustment cell (i.e., in ICILS 2023 within a stratum).

7.3 Calculating weights for the student population

Final student weights were calculated for each eligible and as participating considered student in ICILS 2023. To be considered as “eligible,” the student needs to fit the defined target population, e.g., not older than 17 years, and he or she must be a student of the sampled class and school. To be considered as “participating,” the student worked on the assessment, and the class and the school—the student belongs to—need to be considered as participating, e.g., at least 50 percent of the students in the class participated. For details on participation requirements, see 7.7. The *final student weight* comprises of six weighting components: three base weights and three non-response adjustment weights.

School base weight (WGTFAC1)

The first sampling stage involved selecting schools in each country: the *school base weight* reflects the selection probabilities of this sampling step. As each sampled school is selected for the student, the teacher, and the school study, its selection probability, and therefore its *school base weight*, is the same for the three ICILS 2023 target populations.

When explicit stratification was used, the school samples were selected independently within each explicit stratum h , with $h = 1, \dots, H$. If no explicit strata were formed, the entire country was regarded as being one explicit stratum.

Systematic random samples of schools were drawn in all countries, with a selection probability of school i in stratum h proportional to its size (PPS sampling). The measure of school size M_{hi} was defined by the number of students in the target grade or an adjacent grade. If schools were smaller than the minimum cluster size¹⁷ ($M_{hi} < MCS$), the measure of size M_{hi} was redefined as the average size of all small schools in that stratum. Let M_h the total number of students in stratum h and n_h^s the number of schools to select in that stratum.

The *school base weight* was defined as the inverse of the school's selection probability. For school i in stratum h , the *school base weight* was given by:

$$WGTFAC1_{hi} = \frac{M_h}{n_h^s \times M_{hi}}. \quad (7.1)$$

In census countries or strata, all schools were selected. Here is $WGTFAC1_{hi} = 1$ for all schools in respective countries (Cyprus, Luxembourg, and Malta).

School non-response adjustment (WGTAJ1S) - student population

School base weights for participating schools needed to be adjusted to account for the loss in overall sample size from schools that either refused to participate or had to be removed from the international dataset due to low within-class participation. Adjustments were calculated within non-response groups defined by the explicit strata. A *school non-response adjustment* was calculated for each participating school i within each explicit stratum h as:

17 Minimum cluster size (MCS) is defined as the average class size of all schools in a stratum with two or more classes.



$$WGTADJ1S_{hi} = \frac{n_h^{s,e}}{n_h^{p-std}}, \quad (7.2)$$

where $n_h^{s,e}$ is the number of sampled eligible schools and n_h^{p-std} is the number of participating schools—in the student survey—(whether originally sampled or replacement schools) in explicit stratum h .

The number $n_h^{s,e}$ in this section is not necessarily equal to n_h^s in the preceding section, as $n_h^{s,e}$ was restricted to schools deemed eligible in ICILS 2023. Because of the lapse of one or two years between school sampling and the actual assessment, some selected schools were no longer eligible for participation. This happened if schools had been closed recently, did not have students in the target grade anymore, or had only excluded students enrolled. Ineligible schools such as these were not taken into account when calculating the *school non-response adjustment*.

Class base weight (WGTFAC2S)

The *class base weight* was defined as the inverse of the selection probability of that class. IEA Within-School Sampling Software (WinW3S) was used to manage within-school sampling during the second sampling stage (see [Chapter 6](#) and [Chapter 8](#) for details) in order to conduct a systematic random selection of classes from the target grade. The *class base weight* for a sampled class j in school i in stratum h was calculated as:

$$WGTFAC2S_{hij} = \frac{C_{hi}}{c_{hi}^s}, \quad (7.3)$$

where C_{hi} is the total number of classes with eligible students enrolled in the target grade in this school and c_{hi}^s is the number of sampled classes in this school.

Class non-response adjustment (WGTADJ2S)

In most schools and countries, one class was selected and if this class was considered as not participating (i.e., less than 50% of the students were considered as participating), the respective school would be considered as not-participating and the loss would be covered with the *school non-response adjustment*. In this case *WGTADJ2S* would not be applied to other participating classes.

However, when two or more classes were selected in the second sampling stage and not all of these classes were participating, but at least one, then participating classes of this school were kept and the loss of the non-participating classes was compensated by the *class non-response adjustment*. To minimize the potential bias the adjustment was done over all participating classes within the same explicit stratum and within the same class group if any were defined. Thus $WGTADJ2S_{hij}$ for each participating class j in school i in stratum and class group h was calculated as:

$$WGTADJ2S_{hij} = \frac{c_h^s}{c_h^p}, \quad (7.4)$$

where c_h^s is the total number of sampled classes in stratum and class group h , and c_h^p is the number of participating classes in stratum and class group h .

Student base weight (WGTFAC3S)

Since all eligible students in a sampled class were selected to participate, the *student base weight* was $WGTFAC3S_{hijk} = 1$ for each eligible and participating student.

Student non-response adjustment (WGTADJ3S)

Not all selected, eligible students were willing or able to participate in ICILS 2023. Given the lack of information about absentees, non-participation has to be assumed, for weighting purposes, as being



completely random within schools. This means that participating students represent both participating and non-participating students. The sampling weights of the participating students had to be adjusted accordingly. The *student non-response adjustment* for each participating student k in class j in school i in stratum h was calculated as:

$$WGTADJ3S_{hijk} = \frac{s_{hij}^e}{s_{hij}^p}, \quad (7.5)$$

with s_{hij}^e being the number of eligible students and s_{hij}^p being the number of participating students in the respective class.

In the context of *student non-response adjustment*, students of the target population were regarded as eligible if they had not been excluded due to disabilities or language problems.

Please note that sampled students who did not participate in the survey because they had left the sampled school after class sampling were counted as absent in the sampled school. These students were assumed to remain part of the target population (they moved to a different school but had a zero chance of selection since within-school sampling had already been completed at this point). Excluded students within participating schools carried their weight (i.e., reflecting their proportion in the target population) and this contributed to the overall estimates of exclusion.

Final student weight (TOTWGTS)

The final student weight of student k in class j in school i in stratum h is the product of the student-level-weight components:

$$TOTWGTS_{hijk} =$$

$$WGTFAC1_{hi} \times WGTADJ1_{hi} \times WGTFAC2S_{hij} \times WGTADJ2S_{hij} \times WGTADJ3S_{hijk}. \quad (7.6)$$

Note as $WGTFAC3S = 1$ for all eligible participating students, this factor was removed from the equation.

7.4 Calculating weights for the teacher population

Final teacher weights were calculated for each eligible and as participating considered teacher in ICILS 2023. To be considered as “eligible,” the teacher needs to fit the defined target population, e.g., he or she teaches grade 8 students, and was selected for ICILS 2023. To be considered as “participating,” the teacher worked on the questionnaire, and the school—the teacher belong to—need to be considered as participating, e.g., at least 50 percent of the selected teachers participated, see 7.7. The *final teacher weight* comprises of six weighting components: three base weights and three non-response adjustment weights, see below.

School base weight (WGTFAC1)

Because the same schools were sampled for the student survey and the teacher survey, the *school base weight* of the teacher survey was identical to the *school base weight* of the student survey.

School non-response adjustment (WGTADJ1T)

A *school non-response adjustment* for the teacher survey was calculated in the same way as the *student non-response adjustment*. Given that schools could be regarded as participating in the student survey but not in the teacher survey, and vice versa, the school non-participation adjustment potentially differed between student and teacher data from the same school. To account for non-responding schools in the sample, it was necessary to calculate a *school weight adjustment* for the teacher survey as follows for school i :



$$WGTADJ1T_{hi} = \frac{n_h^{s,e}}{n_h^{p-tch}}, \quad (7.7)$$

where $n_h^{s,e}$ is again the number of sampled eligible schools and n_h^{p-tch} is the number of schools participating (whether originally sampled or replacement schools) in the teacher survey in stratum h .

Teacher base weight (WGTFAC2T)

A systematic random sampling method, carried out via the IEA WinW3S, was used to randomly select teachers in each school. The *teacher base weight* for teacher l was calculated as:

$$WGTFAC2T_{hil} = \frac{T_{hi}}{t_{hi}^s}, \quad (7.8)$$

where T_{hi} is the total number of eligible teachers in school i in stratum h and t_{hi}^s is the number of sampled teachers in school i in stratum h .

In schools with 20 or fewer target grade teachers, all eligible teachers were selected for participation. In these cases this weight factor was equal to one.¹⁸

Teacher non-response adjustment (WGTADJ2T)

Not all teachers were willing or able to participate in the study. Therefore, participating teachers represented both participants and non-participants. Again, the non-response adjustment carried out within a given school assumed, for weighting purposes, that there was a random process underlying teachers' participation.

The *teacher non-response adjustment* was computed for each participating teacher l as:

$$WGTADJ2T_{hil} = \frac{t_{hi}^{s,e}}{t_{hi}^p}, \quad (7.9)$$

where $t_{hi}^{s,e}$ is the number of eligible sampled teachers, and t_{hi}^p is the number of participating teachers in school i in stratum h . Teachers, who left the school after they had been sampled but prior to the data collection, were regarded as out of scope and their weights were not adjusted in these instances.

Teacher multiplicity factor (WGTFAC3T)

Some teachers in ICILS 2023 were teaching at the target grade in more than one school (based on information from the teacher questionnaire) and therefore had a larger selection probability than those teaching at the target grade in only one school. In order to account for this, a "teacher multiplicity factor" was calculated as the inverse of the number of schools in which the teacher was teaching:

$$WGTFAC3T_{hil} = \frac{1}{f_{hil}}, \quad (7.10)$$

where, f_{hil} is the number of schools where teacher l in school i in stratum h was teaching.

Final teacher weight (TOTWGTT)

The final teacher weight for teacher l in school i in stratum h is the product of the five teacher-weight components:

¹⁸ In Luxembourg, the number of teachers to select was increased to 20, thus all eligible teachers were selected in schools with 24 or fewer target grade teachers.



$$TOTWGT_{hil} =$$

$$WGT_{FAC1hi} \times WGT_{ADJ1T_{hi}} \times WGT_{FAC2T_{hil}} \times WGT_{ADJ2T_{hil}} \times WGT_{FAC3T_{hil}}. \quad (7.11)$$

7.5 Calculating weights for the school-level survey

ICILS 2023 was designed as a survey of students and teachers but not as a school survey. However, in order to collect context information at school level, a principal questionnaire and an information and communication technology (ICT) coordinator questionnaire were administered to every participating school. The *final school weights* were calculated and included in the international database in order to allow for analyses of this school-level data. However, results at the school level should be interpreted with some caution as they may be subject to considerable sampling error. As for the student and teacher survey, weights are only calculated for eligible and participating "schools." A school is considered as participating if at least one item was answered by the principal or by the ICT coordinator in the respective questionnaire.

School base weight (WGTFAC1)

This weight component is identical to the *school base weight* of the student survey and the teacher survey (see above).

School weight adjustment (WGTADJ1C)

Schools in which no items were completed in either the principal questionnaire or the ICT coordinator questionnaire were regarded as non-participants in the school survey. In order to account for these non-responding schools, a *school weight adjustment* component was calculated for each participating school as follows:

$$WGT_{ADJ1Chi} = \frac{n_h}{n_h^{p-sch}}, \quad (7.12)$$

where n_h represents the number of eligible sampled schools and n_h^{p-sch} represents the number of schools with completed questionnaires in stratum h (whether originally sampled or replacement schools).

Note that some schools may have been non-participants in the school survey but participated in the student and/or the teacher surveys. Consequently, some schools were regarded as participants in the student and/or teacher survey but as non-participants in the school survey. Some schools may also have completed (at least one of) the school-level questionnaires but were regarded as non-participants in the student and/or teacher surveys. It is very important to keep this in mind when undertaking analysis with data from different data sets. For multivariate analyses using different data sources, the proportion of missing values may accumulate with increasing numbers of variables. Those undertaking secondary analyses should thoroughly monitor the potential loss of information due to missing data across the different sampling units.

Final school weight

The final school weight of school i in stratum h is the product of the two weight components:

$$TOTWGT_{hil} = WGT_{FAC1hi} \times WGT_{ADJ1Chi}. \quad (7.13)$$

7.6 Calculating participation rates

For ICILS 2023, weighted and unweighted participation rates were calculated at student and teacher levels to facilitate the evaluation of data quality and reduce the risk of potential bias due to non-response. In contrast to the weight-adjustments described earlier, participation rates were computed first considering the originally sampled schools only and then considering originally sampled and replacement schools.



Unweighted participation rates in the student survey

Let op denote the set of originally sampled eligible and participating schools, fp the full set of eligible participating schools, including replacement schools, and np the set of sampled eligible but non-participating schools in the student survey. Let n_{op} , n_{fp} , and n_{np} denote the numbers of schools in each of the respective sets. The unweighted school participation rate in the student survey before replacement $UPRS_{schools_{BR}}$ is calculated as:

$$UPRS_{schools_{BR}} = \frac{n_{op}}{n_{fp} + n_{np}}.$$

And the unweighted school participation rate in the student survey after replacement $UPRS_{schools_{AR}}$ is computed as:

$$UPRS_{schools_{AR}} = \frac{n_{fp}}{n_{fp} + n_{np}}.$$

Similarly, let cop denote the set of eligible and participating classes in all participating schools (i.e., including replacement schools), cfp the full set of eligible classes in participating schools, and np the set of eligible but non-participating classes in participating schools. In addition, let n_{cop} , n_{cfp} and n_{cnp} denote the numbers of classes in each of the respective sets. The unweighted class participation rate in the student survey was calculated as:

$$UPRS_{classes} = \frac{n_{cop}}{n_{cfp} + n_{cnp}}.$$

Let sfp represent the set of eligible and participating students in all participating schools, that is, in the schools that constitute fp as the complete set of eligible participating schools. Let snp be the set of eligible but non-participating students in schools that constitute fp , and let n_{sfp} and n_{snp} be the number of students in these two respective groups. The unweighted student response rate is computed as:

$$UPRS_{students} = \frac{n_{sfp}}{n_{sfp} + n_{snp}}.$$

Note that it was not deemed necessary to compute student response rates separately for originally sampled and replacement schools because non-response patterns did not vary between (participating) originally sampled and replacement schools.

The unweighted overall participation rate in the student survey before replacement is then:

$$UPRS_{overall_{BR}} = UPRS_{schools_{BR}} * UPRS_{classes} * UPRS_{students}.$$

And the unweighted overall participation rate in the student survey after replacement is:

$$UPRS_{overall_{AR}} = UPRS_{schools_{AR}} * UPRS_{classes} * UPRS_{students}.$$

Weighted participation rates in the student survey

The weighted school participation rate in the student survey before replacement $WPRS_{schools_{BR}}$ was calculated as the ratio of summations of all participating students k in stratum h and school i and classes j :



$$WPRS_{schools_{BR}} =$$

$$\frac{\sum_h \sum_{i \in op} \sum_j \sum_{k \in sfp} WGT FAC1_{hi} \times WGT FAC2_{S_{hij}} \times WGT ADJ2_{S_{hij}} \times WGT ADJ3_{S_{hijk}}}{\sum_h \sum_{i \in fp} \sum_j \sum_{k \in sfp} WGT FAC1_{hi} \times WGT ADJ1_{S_{hi}} \times WGT FAC2_{S_{hij}} \times WGT ADJ2_{S_{hij}} \times WGT ADJ3_{S_{hijk}}}.$$

Here, the students in the numerator were computed as the sum over the originally sampled participating schools only, whereas the students in the denominator were calculated as the total over all participating schools.

The weighted school participation rate in the student survey after replacement $WPRS_{schools_{AR}}$ is therefore:

$$WPRS_{schools_{AR}} =$$

$$\frac{\sum_h \sum_{i \in fp} \sum_j \sum_{k \in sfp} WGT FAC1_{hi} \times WGT FAC2_{S_{hij}} \times WGT ADJ2_{S_{hij}} \times WGT ADJ3_{S_{hijk}}}{\sum_h \sum_{i \in fp} \sum_j \sum_{k \in sfp} WGT FAC1_{hi} \times WGT ADJ1_{S_{hi}} \times WGT FAC2_{S_{hij}} \times WGT ADJ2_{S_{hij}} \times WGT ADJ3_{S_{hijk}}}.$$

The weighted class participation rate, $WPRS_{classes}$, was calculated as:

$$WPRS_{classes} =$$

$$\frac{\sum_h \sum_{i \in fp} \sum_j \sum_{k \in sfp} WGT FAC1_{hi} \times WGT FAC2_{S_{hij}} \times WGT ADJ3_{S_{hijk}}}{\sum_h \sum_{i \in fp} \sum_j \sum_{k \in sfp} WGT FAC1_{hi} \times WGT FAC2_{S_{hij}} \times WGT ADJ2_{S_{hij}} \times WGT ADJ3_{S_{hijk}}}.$$

The weighted student participation rate $WPRS_{students}$ was computed as follows, taking again replacement schools into account:

$$WPRS_{students} = \frac{\sum_h \sum_{i \in fp} \sum_j \sum_{k \in sfp} WGT FAC1_{hi} \times WGT FAC2_{S_{hij}}}{\sum_h \sum_{i \in fp} \sum_{k \in sfp} WGT FAC1_{hi} \times WGT FAC2_{S_{hij}} \times WGT ADJ3_{S_{hik}}}.$$

The weighted overall participation rate in the student survey before replacement $WPRS_{overall_{BR}}$ is therefore:

$$WPRS_{overall_{BR}} = WPRS_{schools_{BR}} \times WPRS_{classes} \times WPRS_{students}.$$

The weighted overall participation rate in the student survey after replacement $WPRS_{overall_{AR}}$ is:

$$WPRS_{overall_{AR}} = WPRS_{schools_{AR}} \times WPRS_{classes} \times WPRS_{students}.$$

Overview of participation rates in the student survey

Table 7.1 and Table 7.2 display the unweighted and weighted participation rates of all countries in the student survey. Small differences between the two tables indicate different response patterns among strata with disproportional sample allocations. It should be noted that only those schools were treated as participants in the student survey that had at least a participation rate of 50 percent among their sampled students. A school that did not meet this requirement was regarded as a non-participating school for the student data collection. The non-participation of this school had an effect on the school participation rate; however, the students from this school were exempted from the calculation of the student participation rate.



Table 7.1: Unweighted school, class, and student participation rate (PR) - student survey

Country	School PR - unweighted (%)		Class PR - unweighted (%)	Student PR - unweighted (%)	Overall PR - unweighted (%)	
	Before replacement	After replacement			Before replacement	After replacement
Austria	98.1	99.4	100.0	92.9	91.1	92.3
Azerbaijan	94.5	100.0	100.0	92.9	87.8	92.9
Belgium (Flemish)	62.7	88.9	100.0	93.3	58.6	83.0
Bosnia and Herzegovina	97.2	99.1	100.0	90.4	87.9	89.6
Chinese Taipei	97.6	100.0	100.0	95.6	93.4	95.6
Croatia	83.6	97.4	100.0	91.9	76.8	89.5
Cyprus	96.0	96.0	100.0	93.3	89.5	89.5
Czech Republic	100.0	100.0	100.0	94.6	94.6	94.6
Denmark	79.7	95.3	100.0	86.0	68.6	81.9
Finland	100.0	100.0	99.6	92.4	92.1	92.1
France	99.3	99.3	100.0	93.6	92.9	92.9
Germany	86.5	96.5	100.0	90.8	78.5	87.6
Greece	86.6	96.2	100.0	93.7	81.1	90.2
Hungary	97.5	98.7	100.0	93.7	91.3	92.5
Italy	95.4	100.0	100.0	96.5	92.1	96.5
Kazakhstan	97.7	100.0	100.0	96.5	94.3	96.5
Korea, Republic of	76.3	100.0	100.0	95.3	72.7	95.3
Kosovo	99.4	99.4	100.0	94.6	94.0	94.0
Latvia	86.7	95.3	100.0	88.9	77.0	84.7
Luxembourg	100.0	100.0	98.2	90.6	88.9	88.9
Malta	100.0	100.0	100.0	90.8	90.8	90.8
Netherlands	14.2	29.7	96.8	87.6	12.0	25.2
Norway (Grade 9)	95.0	95.6	99.0	86.8	81.6	82.2
Oman	99.5	100.0	100.0	96.0	95.6	96.0
Portugal	88.4	100.0	100.0	94.9	83.9	94.9
Romania	79.5	90.1	100.0	89.4	71.0	80.5
Serbia	98.7	99.4	100.0	93.5	92.2	92.8
Slovak Republic	97.0	100.0	100.0	92.8	90.0	92.8
Slovenia	93.1	96.0	100.0	93.1	86.7	89.4
Spain	99.4	99.6	99.8	90.5	89.8	90.0
Sweden	96.0	98.0	99.4	85.7	81.7	83.4
United States	44.7	65.9	100.0	85.8	38.3	56.6
Uruguay	91.9	96.6	100.0	80.0	73.6	77.4
Benchmarking participant						
North Rhine-Westphalia, Germany	89.0	98.3	100.0	90.5	80.5	88.9

Due to issues with the ICILS main survey data collection in 2023 in Chile, data from Chilean schools were not included in the international report, and are not reported in this table. An additional data collection exercise was subsequently conducted to support the reporting of national data within Chile.



Table 7.2: Weighted school, class, and student participation rate (PR) - student survey

Country	School PR - weighted (%)		Class PR - weighted (%)	Student PR - weighted (%)	Overall PR - weighted (%)	
	Before replacement	After replacement			Before replacement	After replacement
Austria	99.1	99.3	100.0	92.9	92.0	92.3
Azerbaijan	94.5	100.0	100.0	92.6	87.5	92.6
Belgium (Flemish)	64.2	89.1	100.0	92.9	59.7	82.8
Bosnia and Herzegovina	96.1	98.6	100.0	90.7	87.2	89.5
Chinese Taipei	98.1	100.0	100.0	95.6	93.8	95.6
Croatia	82.4	97.2	100.0	92.3	76.1	89.7
Cyprus	96.8	96.8	100.0	93.5	90.5	90.5
Czech Republic	100.0	100.0	100.0	94.5	94.5	94.5
Denmark	79.1	95.3	100.0	86.2	68.2	82.1
Finland	100.0	100.0	99.6	92.3	91.9	91.9
France	99.4	99.4	100.0	93.5	92.9	92.9
Germany	83.6	95.5	100.0	89.1	74.5	85.2
Greece	86.4	95.9	100.0	93.7	81.0	89.9
Hungary	97.5	98.7	100.0	94.0	91.7	92.7
Italy	95.0	100.0	100.0	96.5	91.7	96.5
Kazakhstan	97.7	100.0	100.0	96.4	94.2	96.4
Korea, Republic of	77.2	100.0	100.0	95.3	73.6	95.3
Kosovo	99.3	99.3	100.0	94.8	94.2	94.2
Latvia	87.7	95.3	100.0	89.2	78.2	85.0
Luxembourg	100.0	100.0	94.4	90.8	85.8	85.8
Malta	100.0	100.0	100.0	90.3	90.3	90.3
Netherlands	15.4	30.5	96.1	88.4	13.1	25.9
Norway (Grade 9)	95.2	95.9	99.1	86.7	81.8	82.4
Oman	99.5	100.0	100.0	96.1	95.7	96.1
Portugal	89.0	100.0	100.0	95.0	84.5	95.0
Romania	81.1	90.2	100.0	87.2	70.7	78.7
Serbia	98.6	99.3	100.0	93.1	91.8	92.5
Slovak Republic	97.9	100.0	100.0	92.9	91.0	92.9
Slovenia	93.0	96.0	100.0	93.2	86.7	89.5
Spain	99.3	99.7	100.0	91.0	90.3	90.7
Sweden	96.1	98.1	99.4	86.0	82.1	83.8
United States	44.9	65.1	100.0	86.0	38.6	56.0
Uruguay	90.3	96.8	100.0	80.8	73.0	78.2
Benchmarking participant						
North Rhine-Westphalia, Germany	89.0	98.3	100.0	90.5	80.5	88.9

Due to issues with the ICILS main survey data collection in 2023 in Chile resulting in presented participation rates, data from Chilean schools are not reported in the ICILS 2023 international report. An additional data collection exercise has subsequently been conducted to support the reporting of national data within Chile.



Unweighted participation rates in the teacher survey

The computation of participation rates in the teacher survey follows the same logic as applied in the student survey.

Let op , fp , and np be defined as above, such that the participation status now refers to the teacher survey instead of the student survey, and let n_{op} , n_{fp} , and n_{np} be defined correspondingly. The unweighted school participation rate in the teacher survey before replacement is computed as:

$$UPRT_{schools_{BR}} = \frac{n_{op}}{n_{fp} + n_{np}}.$$

The unweighted school participation rate in the teacher survey after replacement is calculated as:

$$UPRT_{schools_{AR}} = \frac{n_{fp}}{n_{fp} + n_{np}}.$$

Let tfp be the set of eligible and participating teachers in schools that constitute fp , tnp be the set of eligible but nonparticipating teachers in schools that constitute fp , and let n_{tfp} and n_{tnp} be the number of teachers in the respective groups. The unweighted teacher response rate is defined as:

$$UPRT_{teachers} = \frac{n_{tfp}}{n_{tfp} + n_{tnp}}.$$

Note that it was not deemed necessary to compute teacher response rates separately for (participating) originally sampled and replacement schools because the non-response patterns did not vary between sample and replacement schools.

The unweighted overall participation rate in the teacher survey before replacement is computed as:

$$UPRT_{overall_{BR}} = UPRT_{schools_{BR}} \times UPRT_{teachers}.$$

The unweighted overall participation rate in the teacher survey after replacement is calculated as:

$$UPRT_{overall_{AR}} = UPRT_{schools_{[AR]}} \times UPRT_{teachers}.$$

Weighted participation rates in the teacher survey

The weighted school participation rate in the teacher survey before replacement is calculated as:

$$WPRT_{schools_{BR}} = \frac{\sum_h \sum_{i \in op} \sum_{l \in tfp} WGT FAC1_{hi} \times WGT FAC2_{hil} \times WGT ADJ2_{T_{hil}} \times WGT FAC3_{T_{hil}}}{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGT FAC1_{hi} \times WGT ADJ1_{T_{hi}} \times WGT FAC2_{T_{hil}} \times WGT ADJ2_{T_{hil}} \times WGT FAC3_{T_{hil}}}.$$

And the weighted school participation rate in the teacher survey after replacement is calculated as:

$$WPRT_{schools_{AR}} = \frac{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGT FAC1_{hi} \times WGT FAC2_{T_{hil}} \times WGT ADJ2_{T_{hil}} \times WGT FAC3_{T_{hil}}}{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGT FAC1_{hi} \times WGT ADJ1_{T_{hi}} \times WGT FAC2_{T_{hil}} \times WGT ADJ2_{T_{hil}} \times WGT FAC3_{T_{hil}}}.$$

The weighted teacher participation rate is therefore:



$$WPRT_{teachers} = \frac{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGT FAC1_{hi} \times WGT FAC2_{hil} \times WGT FAC3_{hil}}{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGT FAC1_{hi} \times WGT FAC2_{hil} \times WGT ADJ2_{hil} \times WGT FAC3_{hil}}.$$

The weighted overall participation rate in the teacher survey before replacement is calculated as:

$$WPRT_{overall_{BR}} = WPRT_{schools_{BR}} \times WPRT_{teachers}.$$

The weighted overall participation rate in the teacher survey after replacement is computed as:

$$WPRT_{overall_{AR}} = WPRT_{schools_{AR}} \times WPRT_{teachers}.$$

Overview of participation rates in the teacher survey

Table 7.3 and Table 7.4 display the unweighted and weighted participation rates of all countries in the teacher survey. Once more, discrepancies between the two tables indicate differential response patterns between strata with disproportional sample size allocations.

Note that only those schools where at least 50 percent of their sampled teachers had completed the survey were regarded as participants in the teacher survey. A school that did not meet this requirement was regarded as a non-participating school in the teacher survey. The non-participation of this school had an effect on the school participation rate (for the teacher survey), but not on the teacher participation rates as the teachers from this school were not included in the calculation of the teacher participation rates.



Table 7.3: Unweighted school and teacher participation rate (PR) - teacher survey

Country	School PR - unweighted (%)		Teacher PR - unweighted (%)	Overall PR - unweighted (%)	
	Before replacement	After replacement		Before replacement	After replacement
Austria	84.5	85.8	83.3	70.4	71.5
Azerbaijan	84.1	89.6	93.0	78.2	83.4
Belgium (Flemish)	56.2	81.0	77.4	43.5	62.7
Bosnia and Herzegovina	34.9	35.8	81.7	28.5	29.3
Chinese Taipei	96.4	98.8	91.7	88.4	90.6
Croatia	82.9	96.7	88.6	73.4	85.7
Cyprus	89.9	89.9	79.1	71.1	71.1
Czech Republic	100.0	100.0	95.0	95.0	95.0
Denmark	70.3	84.5	74.5	52.4	62.9
Finland	98.1	98.1	90.7	88.9	88.9
France	78.8	78.8	84.5	66.6	66.6
Germany	71.3	78.3	87.7	62.5	68.6
Greece	83.9	92.5	88.3	74.0	81.6
Hungary	89.2	90.4	84.9	75.7	76.8
Italy	94.7	99.3	97.5	92.4	96.9
Kazakhstan	97.7	100.0	99.8	97.5	99.8
Korea, Republic of	75.0	98.7	95.5	71.6	94.2
Kosovo	71.4	71.4	77.5	55.3	55.3
Latvia	80.0	88.7	86.6	69.2	76.7
Luxembourg	43.5	43.5	66.7	29.0	29.0
Malta	73.8	73.8	87.9	64.8	64.8
Netherlands	9.0	19.4	70.9	6.4	13.7
Norway (Grade 9)	86.3	87.5	85.5	73.7	74.8
Oman	98.2	98.6	97.4	95.6	96.1
Portugal	88.4	100.0	87.3	77.2	87.3
Romania	75.5	86.1	91.3	69.0	78.6
Serbia	99.4	100.0	92.7	92.1	92.7
Slovak Republic	95.8	98.8	80.6	77.2	79.6
Slovenia	91.4	94.3	93.8	85.8	88.4
Spain	97.8	98.0	88.7	86.8	87.0
Sweden	93.3	95.3	77.0	71.9	73.4
United States	35.7	52.2	91.1	32.5	47.5
Uruguay	81.9	86.6	77.8	63.7	67.3
Benchmarking participant					
North Rhine- Westphalia, Germany	89.0	98.3	100.0	90.5	80.5

Due to the OSS server hack, some teacher data got lost and couldn't be recovered. This affected the teacher participation rates in Sweden and Norway, i.e., teacher participation rates for these countries are assumed to be higher than reported.

Due to issues with the ICILS main survey data collection in 2023 in Chile resulting in presented participation rates, data from Chilean schools are not reported in the ICILS 2023 international report. An additional data collection exercise has subsequently been conducted to support the reporting of national data within Chile.

7.7 ICILS 2023 standards for sampling participation

It is a significant challenge within countries to achieve full participation (i.e., 100%) in a large-scale assessment and the nature of these challenges vary across countries. Given that one essential purpose of ICILS is to report data at the national level, it is necessary to adjudicate for each country whether the achieved sample is sufficient to warrant scientifically defensible reporting of national estimates. As is customary in IEA studies, ICILS 2023 established guidelines for reporting data for countries with less than full participation. Adjudication of the data was done separately for each participating country and each of the two different ICILS 2023 survey populations. This was carried out in accordance with the recommendations of the sampling referee (Marc Joncas) and in agreement with all members of the ICILS Joint Management Committee.

The first step of the adjudication process was to determine the minimum requirements for within-school participation.

Within-school participation requirements

In general, decreasing response rates entail increasing the risk of getting biased population estimates. Because very little information about non-respondents is available, it was not possible to quantify the risk or bias of estimates due to non-participation in most countries. To overcome this, and in addition to the overall participation rate requirements described below, ICILS 2023 established strict standards for minimum within-school participation: data from schools with a response rate of less than half (50%) of sampled students or teachers, respectively, were discarded from the respective survey. This constraint meant that not every student or teacher who completed a survey instrument was automatically considered as participating and thereby contributing to the computation of population estimates.

The within-school response rate was computed separately for the student survey and the teacher survey. Therefore, a school may count as participating in the student survey but not in the teacher survey or vice versa.

Student survey participation requirements

Students were regarded as respondents if they replied to at least one task in the achievement test. Please note, however, that the amount of partial non-response (i.e., omitted items in questionnaires or tasks that had not been attempted) differed for different items and tasks.

There is evidence that attendance and academic performance tend to be positively correlated (Balfanz & Byrnes, 2012; Hancock et al., 2013). Consequently, the likelihood of biased results may increase as within-school response rates decrease.

Whenever there was evidence that the survey operation procedures in a school had not been conducted following the established ICILS 2023 standards, the corresponding school was regarded as a non-participant. For example, if a school failed to list all eligible students for the selection of a student sample and thus causing a risk of bias due to insufficient coverage, the corresponding school's student data were not included in the final database.

Teacher survey participation requirements

Teachers were regarded as respondents if they replied to at least one item in the teacher questionnaire. But again, as was the situation with respect to the students, the overall amount of partial non-response (i.e., omitted items in the questionnaires) was different for different tasks and items.

It is possible that specific groups of teachers tend to be less likely to participate in a survey. In order to help reduce non-response bias, a school was only regarded as a "participating school" in the teacher survey if at least 50 percent of its sampled teachers participated. If the response rate was lower, teacher data from this school were disregarded and the school was treated as non-participating.



Table 7.4: Weighted school and teacher participation rate (PR) - teacher survey

Country	School PR - weighted (%)		Teacher PR - weighted (%)	Overall PR - weighted (%)	
	Before replacement	After replacement		Before replacement	After replacement
Austria	85.6	86.2	83.9	71.8	72.3
Azerbaijan	83.5	88.9	90.7	75.7	80.6
Belgium (Flemish)	52.6	79.2	77.2	40.6	61.1
Bosnia and Herzegovina	34.2	35.8	80.3	27.5	28.8
Chinese Taipei	95.8	98.7	91.8	88.0	90.6
Croatia	87.1	96.9	89.3	77.8	86.5
Cyprus	91.9	91.9	77.9	71.6	71.6
Czech Republic	100.0	100.0	95.0	95.0	95.0
Denmark	68.7	84.5	74.6	51.2	63.0
Finland	98.2	98.2	91.2	89.5	89.5
France	78.5	78.5	84.2	66.1	66.1
Germany	59.3	68.8	83.7	49.6	57.5
Greece	84.4	92.7	87.5	73.8	81.1
Hungary	89.3	90.4	84.2	75.2	76.1
Italy	95.0	99.3	97.4	92.5	96.7
Kazakhstan	97.9	100.0	99.8	97.7	99.8
Korea, Republic of	77.7	98.8	96.3	74.8	95.1
Kosovo	70.3	70.3	76.3	53.6	53.6
Latvia	80.7	89.6	87.2	70.3	78.1
Luxembourg	43.1	43.1	66.7	28.8	28.8
Malta	73.4	73.4	84.5	62.0	62.0
Netherlands	9.9	19.3	72.0	7.1	13.9
Norway (Grade 9)	87.8	88.4	86.1	75.7	76.1
Oman	97.8	98.0	96.8	94.6	94.8
Portugal	86.3	100.0	88.0	76.0	88.0
Romania	74.7	85.5	91.1	68.0	77.8
Serbia	99.5	100.0	92.1	91.7	92.1
Slovak Republic	94.7	98.9	87.6	82.9	86.7
Slovenia	91.7	94.3	94.0	86.2	88.6
Spain	97.6	98.0	88.4	86.3	86.7
Sweden	92.5	95.5	77.6	71.8	74.1
United States	31.8	50.7	94.0	29.9	47.7
Uruguay	79.6	87.7	78.1	62.2	68.5
Benchmarking participant					
North Rhine-Westphalia, Germany	89.0	98.3	100.0	90.5	80.5

Due to the OSS server hack, some teacher data got lost and couldn't be recovered. This affected the teacher participation rates in Sweden and Norway, i.e., teacher participation rates for these countries are assumed to be higher than reported.

Due to issues with the ICILS main survey data collection in 2023 in Chile resulting in presented participation rates, data from Chilean schools are not reported in the ICILS 2023 international report. An additional data collection exercise has subsequently been conducted to support the reporting of national data within Chile.

If a school failed to follow the survey operation procedures properly, it was classified as a non-participating school. For example, if a school failed to list all eligible teachers for the teacher sample selection, or if the standard teacher selection procedures had not been followed, then teacher data from this particular school were not included in the final database.

Participation requirement categories

Three categories for sampling participation were defined the same for the student and the teacher population:

- Countries grouped in Category 1 met the ICILS 2023 sampling participation requirements.
- Countries in Category 2 met these requirements only after the inclusion of replacement schools.
- Countries in Category 3 failed to meet the ICILS 2023 sampling participation requirements.

The results from ICILS 2023 show that high response rates in the teacher survey were often harder to achieve than in the student survey. However, there is no statistical justification to apply different sampling participation standards to the two surveys. Since non-response holds a high potential for bias in both parts of the study, the participation requirements in the teacher survey were identical to those in the student survey. No participation requirements were determined for the reporting of school-level data, however, the participation rate in the school survey was above 85 percent for all countries that were placed in Category 1 and Category 2 for the student survey.

The three categories for sampling participation were defined according to the criteria presented in [Figure 7.1](#).

Figure 7.1: Participation categories in ICILS

Category 1: Satisfactory sampling participation rate without the use of replacement schools.

In order to be placed in this category, a country has to have:

- An unweighted school response rate without replacement of at least 85 percent (after rounding to the nearest whole percent) AND an unweighted overall student/teacher response rate (after rounding) of at least 85 percent
OR
- A weighted school response rate without replacement of at least 85 percent (after rounding to the nearest whole percent) AND a weighted overall student/teacher response rate (after rounding) of at least 85 percent OR
- The product of the (unrounded) weighted school response rate without replacement and the (unrounded) weighted overall student/teacher response rate of at least 75 percent (after rounding to the nearest whole percent).

Category 2: Satisfactory sampling participation rate only when replacement schools were included.

A country is placed in this category if: It fails to meet the requirements for Category 1 but has either an unweighted or weighted school response rate without replacement of at least 50 percent (after rounding to the nearest percent) AND HAS EITHER

- An unweighted school response rate with replacement of at least 85 percent (after rounding to the nearest whole percent) AND an unweighted overall student/teacher response rate (after rounding) of at least 85 percent
OR
- A weighted school response rate with replacement of at least 85 percent (after rounding to nearest whole percent) AND a weighted overall student/teacher response rate (after rounding) of at least 85 percent
OR
- The product of the (unrounded) weighted school response rate with replacement and the (unrounded) weighted overall student/teacher response rate of at least 75 percent (after rounding to the nearest whole percent).

Category 3: Unacceptable sampling response rate even when replacement schools are included.

Countries that can provide documentation to show that they complied with ICILS sampling procedures, but do not meet the requirements for Category 1 or Category 2 are placed in Category 3.

7.8 Reporting data

The ICILS 2023 research team considered it necessary to make readers of the international reports aware of the increased potential for bias, regardless of whether such a bias was actually introduced. Based on their respective sample participation categories, national survey results were reported on as follows:

- Category 1: Countries in this category appear in the tables and figures in the international reports without annotation.
- Category 2: Countries in this category are annotated in the tables and figures in the international reports.
- Category 3: Countries in this category appear in a separate section of the tables and figures.

Table 7.5 shows the used annotation symbols and their meaning for the student and the teacher survey. Table 7.6 lists for each country its reporting category and the corresponding annotation for the student survey, while Table 7.7 lists the same for the teacher survey.

For the student survey, 26 countries and one benchmarking participant were reported as meeting sampling participation requirements in Category 1. Of these, 13 countries got an annotation due to exceeding five percent exclusion limit. Five countries were in Category 2, thereof two with additional annotations due to exclusion rates larger than five percent. One country was reported as Category 3; and two countries were not reported: Chile conducted an additional national data collection, completing the ICILS 2023 data and requested to be not reported, while the participation rates from the Netherlands were too low to be reported. Compared to earlier cycles, the number of countries excluding between five percent and ten percent of their student population has increased.

For the teacher survey, 14 countries and one benchmarking participant had high participation rated as Category 1. Three countries were reported in Category 2, and 12 countries were reported as a Category 3 country. Five countries were not reported due to low participation rates. Please note that additional annotations were needed for not following the sampling procedures, and for excluding more than five percent of students on school-level. For the latter case, it is to be expected that also more than five percent of the teacher would then not be covered, see Table 7.7.

Table 7.5: Used Annotations for the student and teacher survey

Annotation symbol	Annotation note
Annotation note for Participation Rates	
†	Participation Category 2: Met guidelines for sampling participation rates only after replacement schools were included.
‡	Participation Category 3: Does not meet guideline for sampling participation rate, but achieved at least 50 percent overall sampling participation rate.
Annotation note: Coverage & Exclusions	
1	National defined population covers 90 percent to 95 percent of the national target population.
3	National defined population covers 61 percent of the national target population.
Other annotations	
2	Country surveyed target grade in the first half of the school year.
Ψ	Sampling procedures for teacher listing were not followed and any conclusions about teacher analysis is subject to caution (possible biases in both point estimates and standard errors)—Teacher study only.



Table 7.6: Reporting category of each country for the student survey

Country	Reporting category	Annotation symbol(s)	Annotation note: Participation Rates	Annotation note: Coverage & Exclusions	Other annotation note
Austria	1	1		National defined population covers 90% to 95% of the national target population.	
Azerbaijan	1				
Belgium (Flemish)	2	†	Met guidelines for sampling participation rates only after replacement schools were included.		
Bosnia and Herzegovina	1	3		National defined population covers 61% of the national target population.	
Chile	Not reported		Not reported as requested by country, due to additional national data collection in school year 2023/2024. Data would be reported in Category 3.		
Chinese Taipei	1				
Croatia	1	1		National defined population covers 90% to 95% of the national target population.	
Cyprus	1				
Czech Republic	1	1		National defined population covers 90% to 95% of the national target population.	
Denmark	2	†, 1	Met guidelines for sampling participation rates only after replacement schools were included.	National defined population covers 90% to 95% of the national target population.	
Finland	1				
France	1				
Germany	1				
Greece	1				
Hungary*	1				
Italy	1				
Kazakhstan	1	1		National defined population covers 90% to 95% of the national target population.	
Korea, Republic of	2	†	Met guidelines for sampling participation rates only after replacement schools were included.		
Kosovo	1	1		National defined population covers 90% to 95% of the national target population.	

*Hungary did not apply double coding for occupations, i.e., data processing was not fully applied to the rules—not annotated.



Table 7.6: Reporting category of each country for the student survey (cont'd)

Country	Reporting category	Annotation symbol	Annotation note: Participation Rates	Annotation note: Coverage & Exclusions	Other annotation note
Latvia	1	1		National defined population covers 90% to 95% of the national target population.	
Luxembourg	1				
Malta	1				
Netherlands	Not reported		Does not meet guideline for sampling participation rate for reporting.		
Norway (Grade 9)	1	1		National defined population covers 90% to 95% of the national target population.	
Oman	1				
Portugal	1	1		National defined population covers 90% to 95% of the national target population.	
Romania	2	†, 1, 2	Met guidelines for sampling participation rates only after replacement schools were included.	National defined population covers 90% to 95% of the national target population.	Country surveyed target grade in the first half of the school year.
Serbia	1	1		National defined population covers 90% to 95% of the national target population.	
Slovak Republic	1				
Slovenia	1	1		National defined population covers 90% to 95% of the national target population.	
Spain	1	1		National defined population covers 90% to 95% of the national target population.	
Sweden	1	1		National defined population covers 90% to 95% of the national target population.	
United States	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Uruguay	2	†	Met guidelines for sampling participation rates only after replacement schools were included.		
Benchmark participant					
North Rhine-Westphalia, Germany	1				



Table 7.7: Reporting category of each country for the teacher survey

Country	Reporting category	Annotation symbol	Annotation note: Participation Rates	Annotation note: Coverage & Exclusions	Other annotation note
Austria	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Azerbaijan	1				
Belgium (Flemish)	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Bosnia and Herzegovina	Not reported		Does not meet guideline for sampling participation rate for reporting.		
Chile	Not reported		Not reported as requested by country, due to additional national data collection in school year 2023/2024. Data does not meet guideline for sampling participation rate for reporting.		
Chinese Taipei	1				
Croatia	1				
Cyprus	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Czech Republic	1				
Denmark	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Finland	1				
France	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Germany	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		



Table 7.7: Reporting category of each country for the teacher survey (cont'd)

Country	Reporting category	Annotation symbol	Annotation note: Participation Rates	Annotation note: Coverage & Exclusions	Other annotation note
Greece*	2	†	Met guidelines for sampling participation rates only after replacement schools were included.		
Hungary	1				
Italy	1				
Kazakhstan	1				
Korea, Republic of	1				
Kosovo	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Latvia	2	†, 1	Met guidelines for sampling participation rates only after replacement schools were included.	National defined population covers 90% to 95% of the national target population.	
Luxembourg	Not reported		Does not meet guideline for sampling participation rate for reporting.		
Malta	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Netherlands	Not reported		Does not meet guideline for sampling participation rate for reporting.		
Norway (Grade 9)***	1				
Oman**	3				Sampling procedures for teacher listing were not followed and any conclusions about teacher analysis is subject to caution (possible biases in both point estimates and standard errors).

*Greece included also teachers who joined within the first two month after the beginning of the school year as this happens quite frequently. Not annotated.

**Participation rates were of Category 1, however not following sampling procedures for teacher listing might have introduced bias into the data.

***Due to the OSS server hack, some teacher data got lost and couldn't be recovered. This affected the teacher participation rates in Sweden and Norway, i.e., teacher participation rates for these countries are assumed to be higher than reported.



Table 7.7: Reporting category of each country for the teacher survey (cont'd)

Country	Reporting category	Annotation symbol	Annotation note: Participation Rates	Annotation note: Coverage & Exclusions	Other annotation note
Portugal	1	1		National defined population covers 90% to 95% of the national target population.	
Romania	2	†, 2	Met guidelines for sampling participation rates only after replacement schools were included.		Country surveyed target grade in the first half of the school year.
Serbia	1				
Slovak Republic**	3				Sampling procedures for teacher listing were not followed and any conclusions about teacher analysis is subject to caution (possible biases in both point estimates and standard errors).
Slovenia	1				
Spain	1				
Sweden***	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
United States	Not reported		Does not meet guideline for sampling participation rate for reporting.		
Uruguay	3	‡	Does not meet guideline for sampling participation rate, but achieved at least 50% overall sampling participation rate.		
Benchmark participant					
North Rhine-Westphalia, Germany	1				

**Participation rates were of Category 1, however not following sampling procedures for teacher listing might have introduced bias into the data.

***Due to the OSS server hack, some teacher data got lost and couldn't be recovered. This affected the teacher participation rates in Sweden and Norway, i.e., teacher participation rates for these countries are assumed to be higher than reported.



References

- Balfanz, R., & Byrnes, V. (2012). *Chronic absenteeism: Summarizing what we know from nationally available data*. University Center for Social Organization of Schools.
- Franklin, S., & Walker, C. (Eds.). (2003). *Survey methods and practices*. Statistics Canada. <https://www150.statcan.gc.ca/n1/pub/12-587-x/12-587-x2003001-eng.pdf>
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Wiley.
- Hancock, K. J., Shepherd, C. C. J., Lawrence, D., & Zubrick, S. R. (2013). Student attendance and educational outcomes: Every day counts. *Department of Education, Employment and Workplace Relations*. <https://doi.org/10.13140/2.1.4956.6728>
- Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.
- Meinck, S., & Cortes, D. (2015). Sampling weights, nonresponse adjustments, and participation rates. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *IEA International Computer and Information Literacy Study 2013 technical report* (pp. 87–112). International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2013-technical-report>
- Meinck, S. (2015). Computing sampling weights in large-scale assessments in education. *Survey methods: Insights from the field*. <https://doi.org/10.13094/SMIF-2015-00004>
- Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*. CRC Press.
- Tieck, S. (2020). Sample weights, non-response adjustments, and participation weights. In J. Fraillon, J. Ainley, W. Schulz, T. Friedman, & D. Duckworth (Eds.), *IEA International Computer and Information Literacy Study 2018 technical report* (pp. 79–93). International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>

Chapter 8:

ICILS field operations

Sebastian Meyer

8.1 Introduction

Successful administration of ICILS 2023 depended heavily on the contributions of the study's national research coordinators (NRCs) and national center staff. To assist the national centers and promote consistency in the study administration across countries, the ICILS International Study Center (ISC) at IEA developed internationally standardized field operations procedures. The ICILS operations procedures were designed to be flexible enough to meet the needs of individual participants whilst maintaining IEA's high quality standards.

The ICILS operations procedures were developed with reference to those used in ICILS 2018 and in other IEA studies and were tailored to suit the specific requirements of ICILS 2023. All national centers received guidelines on survey operations procedures for each stage of the study. These guidelines provided information about contacting schools, listing and sampling students, preparing materials for data collection, administering the assessment, scoring and coding the assessment, creating the data files, and implementing quality control procedures at various stages.

8.2 Field operations personnel

The role of the national research coordinators and their centers

Appointing an NRC was one of the first steps for all countries or education systems participating in ICILS 2023. The NRC acted as the main coordinator and point of contact for ICILS 2023 at both the national and international level. They were in charge of the overall implementation of the study at the national level. They also, where necessary, implemented and adapted internationally agreed-upon procedures for the national context under the guidance of the ISC and national experts.

The role of school coordinators and test administrators

Two roles were essential to facilitate the successful national administration of ICILS 2023 within countries: the school coordinator and the test administrator. Their work involved preparing for the test administration in schools and carrying out the data collection in a standardized way.

National centers identified and trained school coordinators for all participating schools, most often in coordination with school principals. The school coordinator could be a teacher or other staff member in the school. In some cases, national centers appointed external individuals as school coordinators. The coordinators' responsibilities included the following major tasks:

- Listing target grade classes and identifying teachers belonging to the target population to allow the national center to perform within-school sampling;
- Arranging the date(s) and modalities of the test administration, in particular the delivery method of the student test, with the national center;
- Distributing related materials needed for test administration and making sure they were kept in a secure place and confidential at all times;
- Working with the school principal, the test administrator, and the affected teachers to plan and administer the student test;
- Confirming with the national center the status and eligibility of students qualified to use some form of accommodation to complete the ICILS instruments; and



- Ensuring that the test administrators return all testing materials after the testing session.

The test administrators were mainly responsible for running the assessment sessions in schools where students completed the online test and questionnaire. Test administrators were employed either by the national center or directly by the schools, but could not be a teacher of any of the sampled students. A training session was run by the national center centrally or by the schools to make sure that the test administrators were adequately prepared to run the assessment sessions. In some cases, the school coordinators also served as the test administrators.

8.3 Field operations resources

Manuals and documentation

The international study team released the ICILS 2023 Survey Operations Procedures (SOP) manuals to the NRCs in five units. Each unit was accompanied by additional materials such as manuals for use in schools, software manuals and packages, and informational webinars. The materials were organized and distributed chronologically according to the stages of the study.

The five units were as follows:

- **Unit 1: Sampling Schools** specified the actions and procedures required to develop a national sampling plan in compliance with the international ICILS 2023 sample design.
- **Unit 2: Working with Schools** contained information about how to work with schools to plan the successful administration of the ICILS 2023 instruments.

Accompanying manuals:

- The *Compatibility Check and School Computer Resources Survey: Instructions for NRCs* described the tasks for testing computers for the computer-based ICILS assessment administration, including some technical troubleshooting scenarios.
- The *Preparing Computers for ICILS Manual* described the tasks for preparing computers for the computer-based ICILS assessment administration.

- **Unit 3: Preparing National Study Instruments** described the processes involved in preparing the ICILS 2023 instruments for production and use in countries.

Accompanying manuals:

- The *Assessment Master Manual* provided technical information on how to use the RM Australia Assessment Master throughout the process of preparing national instruments (translation and translation verification).
- The *IEA Online SurveySystem Manual* provided technical information on how to use the IEA Online SurveySystem for setting up the online adult questionnaires.

- **Unit 4: Data Collection and Quality Monitoring Procedures** dealt with the processes involved in preparing for, supporting, and monitoring ICILS 2023 data collection in schools.

Accompanying manuals:

- The *School Coordinator Manual*, described the role and responsibilities of the school coordinator, the main contact person within each participating school.
- The *Test Administrator Manual*, described the role and responsibilities of the test administrator, whose work included administration of the student assessment.

- **Unit 5.1: Monitoring Participation and Data Submission** provided guidelines on tracking the data collection progress and described the data uploading of student assessment data (when the USB delivery method was used).



- **Unit 5.2: Scoring and Parental Occupation Coding** provided guidelines on scoring student responses, and coding parental occupations.

Accompanying manuals:

- The *Scoring Guides*, provided detailed and explicit guidelines on how to score each constructed-response item.

The international versions of some of the manuals could be adapted to account for the national context. In addition, manuals distributed to national staff could be translated to the national language(s) within the country.

Software

To digitally deliver the student tests and questionnaire, ICILS 2023 made use of the Assessment Master (AM), an assessment platform owned by RM Results (see [Chapter 3](#) for further details). Prior to data collection, the ISC provided countries with compatibility check software for checking the suitability of computer devices for ICILS 2023 for the intended delivery mode. During data collection, national center staff could use AM's Manager module to monitor the data collection progress.

In addition to the computer-based delivery software and consistent with the goal of automating and streamlining procedures wherever possible, the ISC provided NRCs with custom-built software packages to assist with sampling and data collection:

- IEA Windows Within-School Sampling Software (WinW3S): This enabled the ICILS 2023 national centers to randomly select classes and teachers in each sampled school in agreement with sample design specifications and mandatory sampling algorithms. National centers used the software to track school, teacher, and student information, prepare the survey tracking forms, and assign test instruments to students.
- IEA Online Survey System (OSS): was used as a delivery system for the online adult questionnaires (for teachers, ICT-coordinators and school principals). It included a designer module for entering verified translations and implementing agreed upon national adaptations to the international version of the questionnaires and data monitoring module that allowed the tracking of principal, ICT coordinator, and teacher responses flowing into the system.

In addition to providing software and manuals, the ISC conducted field operations training seminars prior to both the field trial and the main survey. These seminars were designed to train national center staff on all procedures and to familiarize them with the supporting software, namely IEA WinW3S and IEA OSS. The ISC also offered a webinar immediately after the release of the field trial and main survey instruments to explain translation procedures within the AM Designer module.

8.4 Field operations processes

Linking students and teachers to schools

The ISC established a system for assigning hierarchical identification codes (IDs) to uniquely identify and track sampled schools, teachers, and students. [Table 8.1](#) presents the hierarchical identification system.

Each sampled student was assigned an eight-digit ID number unique within their country. This ID consisted of a four-digit school identifier, followed by a two-digit class identifier within the school, and a two-digit student identifier within the class.

Similarly, each sampled teacher within a sampled school was assigned a teacher ID number, consisting of the four-digit school identifier followed by a two-digit teacher identifier unique within the school.

Table 8.1: Hierarchical identification codes

Unit	ID components	ID structure	Numeric example
School (Principal and ICT-Coordinator)	School (C)	CCCC	1001
Class	School (C), Class ID (K)	CCCCKK	100101
Student	School (C), Class ID (K), Student (S)	CCCCKKSS	10010101
Teacher	School (C), Teacher (T)	CCCCTT	100101

Activities for working with schools

In ICILS 2023, the within-school sampling process and the assessment administration required close cooperation between the national centers and representatives from the schools (i.e., the school coordinators and test administrators as described previously). [Figure 8.1](#) presents the major activities the national centers conducted when working with schools to list and sample students and teachers, track respondents, prepare for test administration, and carry out the assessment.

Contacting schools and within-school sampling procedures

Once national centers had obtained a list of the schools sampled for ICILS 2023, it was important for the success of the study to establish good working relationships with the selected schools (for more information on sampling procedures, please refer to [Chapter 6](#) of this report). The national centers were responsible for contacting the schools and encouraging them to participate in the assessment, a process that often involved obtaining support from national or regional educational authorities or other stakeholders, depending on the national context.

As described earlier, national centers identified and trained school coordinators for all participating schools. School coordinators were responsible for providing all required information about their respective schools and coordinating the date, time, and place of the student assessment. School coordinators were also responsible for arranging modalities of the test administration with the national center, for example, regarding the use of school or externally provided computers. This work required them to complete the school computer resources survey, run compatibility checks locally, and send results to the national study center.

National centers sent a class listing form to each school coordinator and asked them to provide information on all target-grade classes in the school, including the number of students in each class. The national centers used this information to sample classes within the schools.

For each sampled class, the national centers created a student listing form which was sent to the schools coordinator. School coordinators were required to list all students within the sampled class along with selected details about these students, such as their names (if country regulations allowed names to be provided to the national centers), birth month and year, gender, exclusion status, the assessment language of the student (in case the national center provided different language versions of the student instruments), and if any special accommodations for the student test sessions were necessary.

National centers also sent a teacher listing form to each school coordinator and asked them to provide information on all the eligible target-grade teachers within the school. The school coordinators listed the eligible target-grade teachers and provided details about these teachers, such as their names (if country regulations allowed names to be provided to the national center), birth month and year, and gender. The national centers used the collected information to sample teachers within the schools.

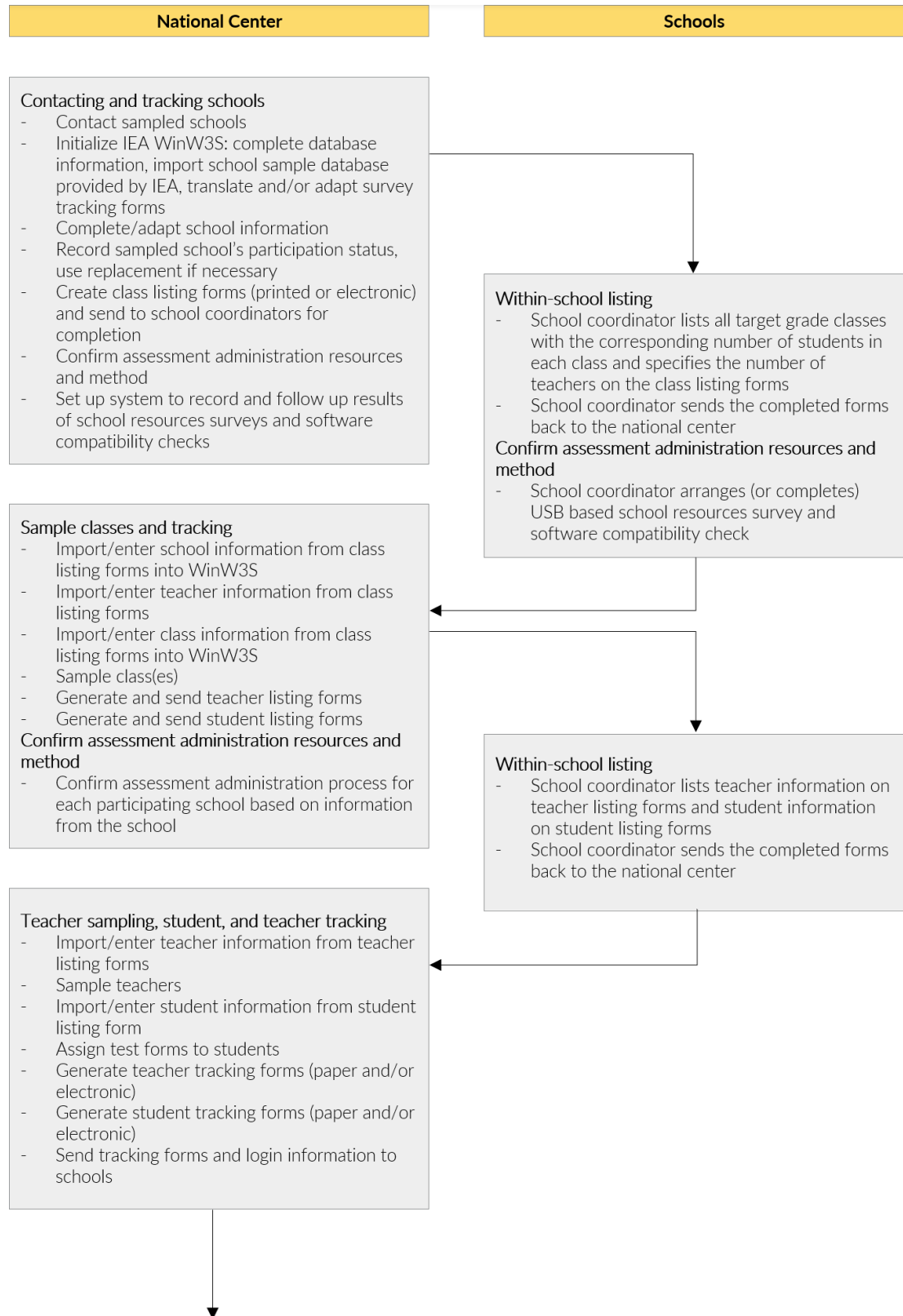


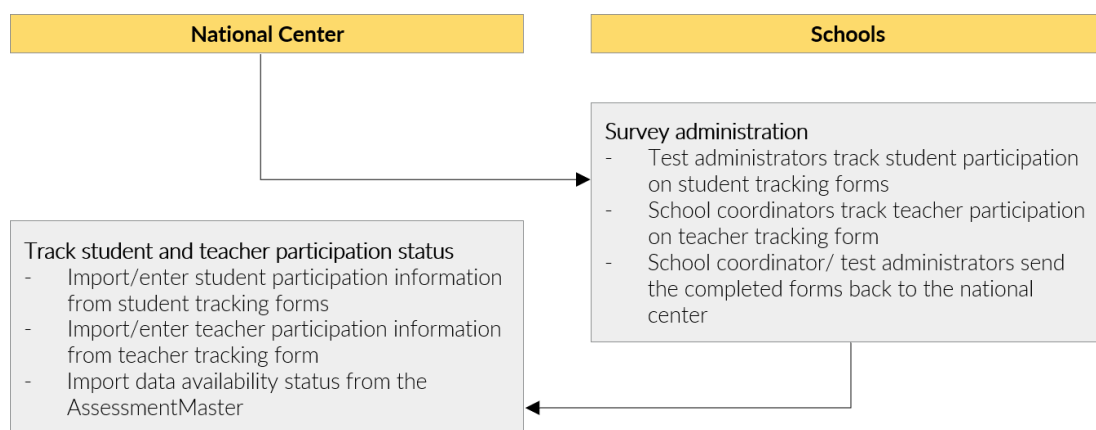
School coordinators were also responsible for obtaining parental permission as necessary, liaising with the test administrator to coordinate the test session, distributing information about teacher, school, and ICT coordinator questionnaires, and coordinating completion of the student tracking forms and teacher tracking forms. School coordinators also ensured that assessment materials were received, kept secure at all times, and returned to the national center after the administration.

The following [Figure 8.1](#) illustrates the exchange of information between the national centers and the participating schools.



Figure 8.1: Procedures for working with schools to prepare for test administration





Preparing the computer-based test delivery at schools

Because ICILS 2023 was a computer-based assessment, it was necessary to test the computer resources available at participating schools to ascertain whether the school computer resources could be used to deliver the assessment. Two delivery methods were recommended for the student test:

- Internet (browser-based) method; and
- USB flash drive method.

Countries were asked to select both a primary and a backup delivery method, with no obligation to use only one method exclusively, although doing so was recommended for organizational purposes. Countries were also given the possibility to arrange for portable computers to be transported to schools that use USB flash drives to deliver the *ICILS Student Software*. See [Chapter 3](#) for details of the system requirements for the different delivery methods.

The compatibility check and school computer resources survey were administered in order to answer three questions: (i) if school computers could be used for the testing or if schools would need to be provided with computers able to do this task; (ii) if, in those cases where the school computers were intended to be used via online delivery, the school provided sufficient bandwidth to deliver the testing through web browser; and (iii) if, in those cases where the school computers were intended to be used for USB flash drive delivery, special arrangements would be needed (e.g., altering the configuration of computers) for the student test to run correctly.

Schools were requested to run the compatibility check and complete the computer resources survey far enough in advance of the ICILS testing to allow for any problem solving and modifications to computer configurations in preparation for the ICILS testing but sufficiently close to the ICILS testing day to minimize the risk of computer configurations changing greatly between the time that they were checked and the ICILS testing date.

In schools where the student test was planned to be delivered using the internet method, school coordinators were asked to access the ICILS browser compatibility check website to verify whether the browser and screen resolution were suitable on each testing computer. They were also required to use the internet latency check website to measure the response time between at least one testing computer and the Amazon Web Services (AWS) regions.

In schools where the student test was planned to be delivered using the USB method, school coordinators were instructed to run a program from a USB flash drive to confirm whether the minimum computer requirements were met. This program needed to be executed on all computers designated for testing.



The school resources survey specifically requested the results of the compatibility checks. The completed surveys were required to be sent back to the national center, where they were evaluated to assess the specific situation in each participating school. If the evaluation identified any risks that could compromise the feasibility of the planned delivery method, the national center could implement targeted measures to mitigate those risks. National centers in non-English-speaking countries were also asked to translate the school computer resources survey questions.

Once the technical requirements were met, the national centers sent the following items to each school: tracking forms, labels, manuals, and, where applicable, USB sticks for each student listed on the student tracking form, along with three extra sticks.

Administering the assessment at schools

Preparing assessment materials and distributing them to the participating schools required careful organization and planning by national centers. The assessment materials were provided to the school coordinators prior to testing, providing ample time for the school coordinators to confirm the receipt and correctness of the materials.

For all survey instruments, individual login information was provided to ensure confidentiality. For the principal, ICT coordinator, and teacher questionnaires, the national centers sent the login information, along with general information on how to access the online questionnaire, to the school coordinator in the form of “cover letters.” The school coordinator delivered the principal, ICT coordinator, and teacher cover letters to the designated individuals. Student login information was provided to the school coordinator with the student tracking forms and kept in a secure room until they were provided to the test administrators.

In accordance with the international guidelines and requirements as well as local conditions, national centers assigned a test administrator to each school as described earlier. Their responsibilities included: running a pretest administration on the day of testing in order to confirm that the student computers were prepared for the test; distributing materials to the appropriate students; logging in and initializing the test on the computers (either via the USB sticks provided by the national centers or the internet method); leading students through the assessment; and, accurately timing the sessions.

For each sampled student, the student tracking forms indicated the assigned student instrument which consisted of either two or three parts depending on whether or not the computational thinking (CT) modules were administered within the country. Each sampled student was assigned two out of seven computer and information literacy (CIL) modules from a completely balanced design so that each module was assigned to an approximately equal number of students. Each student also was asked to complete a questionnaire. If a country opted for administering the CT modules, each student was also asked to complete two out of four additional CT modules which also were assigned from a balanced design.

During the administration of the assessment, administrators were required to provide a range of instructions to students. For some parts of the assessment, test administrators were asked to read instructions to the students as provided to them in the *Test Administrator Manual*. Administrators had to read the text to the students exactly as it appeared in the script. In some other parts of the assessment test, administrators were required to read instructions from a script but had the option of modifying or adapting it to best suit a given situation.

In these instances, it was essential that the exact contents and meaning of each of the scripts was conveyed to each set of students. The only instances in which test administrators could use their own words was when the *Test Administrator Manual* did not include a script for the instructions, for example, when the manual explicitly advised administrators that they could answer any questions or points of clarification. Please see [Chapter 9](#) for data on how well test administrators adhered to the script.

The time allotted for each part of the student testing and questionnaire administration was standardized across countries. In all countries, target-grade students were allowed 30 minutes to complete each of the two modules (60 minutes in total). Students who completed the assessment before the allotted time was over were allowed to review answers or read quietly, but were not allowed to leave the session. Students were given around 25 minutes to complete the student questionnaire and were allowed to continue if they needed additional time. Test administrators were required to document the starting and ending time of each part of the assessment administration on the test administration form.

In countries administering the two CT modules, after the regular CIL test and student questionnaire sessions, the test session was extended by an additional 50 minutes (25 minutes for each of the two CT test modules). [Table 8.2](#) details the time allotted to the different parts of the student assessment.

Once the administration was completed, the school coordinators were responsible for collecting and returning all materials to their respective national center.

Table 8.2: Timing of the ICILS assessment

Activities	Length
Preparation of students, reading of instructions, and administering the tutorial	20 minutes (approx.)
Administering the CIL student assessment—first module	30 minutes (exact)
Short break	5 minutes (max.)
Administering the CIL student assessment—second module	30 minutes (exact)
Short break	5 minutes (max.)
Administering the student questionnaire	20 minutes (approx.)
Longer break (CT only)	Between 15 and 45 minutes
Administering the CT student assessment—first module (CT only)	25 minutes (exact)
Administering the CT student assessment—second module (CT only)	25 minutes (exact)
Collecting the assessment materials and ending the session	5 minutes (approx.)
TOTAL (CIL only)	2 hours (approx.)
TOTAL (including CT)	3.5 hours (approx.)

Online data collection of school principal, ICT coordinator, and teacher questionnaires

In ICILS 2023, it was mandatory to deliver all adult questionnaires online. After the principal, ICT coordinator, and teacher questionnaires had gone through the instrument preparation and verification processes, they were prepared for delivery online using the IEA Online Survey System (IEA OSS) software. These procedures and software are described in more detail in [Chapter 5](#) of this report.

The electronic versions of the ICILS 2023 principal, ICT coordinator, and teacher questionnaires could only be completed via the internet. Accordingly, the design ensured that respondents needed only an internet connection and a standard internet browser. No additional software or particular operating system was required. Respondents were not allowed to use other delivery options, such as sending PDF documents via email or printing out the online questionnaires and mailing them to the national center.

During the administration period, respondents could log in and out as many times as they needed and could resume answering the questionnaire at the question they had last responded to in their previous



session. Answers were automatically saved whenever respondents moved to another question, and respondents could change any answer at any time before completing the questionnaire. During the administration, the national center was available for support; the center, in turn, could contact the ISC if unable to solve a problem locally.

Responses to the online questionnaires were not made mandatory, evaluated, or enforced in detail (e.g., using hard validations). Instead, some questions used soft validation, such as respondents being asked to give numerical responses to questions that had a minimum and maximum value, for example, the total number of students enrolled in a school. In some instances, respondents' answers to this type of question led to the response being updated according to the individual respondent's entries, even if that response was outside the minimum or maximum value, but with the caveat that the response still needed to be within the specified width.

To reduce response burden and complexity, the online survey automatically skipped questions not applicable to the respondent. Rather than presenting multiple questions per page, the online questionnaire proceeded question by question. While vertical scrolling was required for a few questions, particularly the longer questions with multiple "yes/no" or Likert-type items, horizontal scrolling was not. Because respondents could easily estimate through visual cues the length and burden of a paper questionnaire, the online questionnaires attempted to offer this feature through progress counters and a "table of contents" that listed each question and its response status. Multiple-choice questions were implemented with standard HTML radio buttons.

Because the national centers were able to monitor the responses to the online questionnaires in real-time, they could send reminders to those schools where people had not responded in the expected period of time. Typically, in these cases, the centers asked the school coordinators to follow up with those individuals who had not responded. It was also possible to provide school coordinators with access to the monitoring tool. In that case national centers were asked to send a list including the number of school coordinators and the school IDs each school coordinator was overseeing to IEA. That way, school coordinators were only seeing the progress of the schools they are responsible for.

Scoring the assessment

The success of assessments containing constructed-response items and large tasks depends on the degree to which student responses are scored reliably. Nineteen of the ICILS 2023 CIL assessment tasks were constructed-response tasks, and seven large tasks were scored against a total of 60 criteria. One of the large-task criteria was automatically scored by the Assessment Master Examiner module. Human scorers reviewed the automatically generated suggested score and could either accept or modify the score. Of the 55 ICILS 2023 CIL tasks, 26 were scored by human scorers. There were 31 CT tasks for ICILS 2023. Three of these tasks involved human scoring. All human scoring was conducted within the Assessment Master Marker module. It was critical to the quality of the ICILS 2023 results that these tasks were scored in a reliable manner. Reliability was accomplished by providing national centers with explicit scoring guides, extensive training of scoring staff, and continuous monitoring of the quality of the work during scoring procedures.

During the scoring training, which was conducted at the international level, national center staff members learned how to score the constructed-response items and to use the scoring criteria for the large-task items in the ICILS 2023 assessment. Scoring training took place before both the field trial and the main survey. The training that took place prior to the field trial provided the participants with their first opportunity to give extensive feedback on the scoring guides, which were then revised on the basis of this feedback. The training conducted before the main survey enabled national center staff to give additional feedback on the scoring guides, with that feedback based on their experiences of scoring the field-trial items. The content of the scoring guides for the five ICILS trend modules (four CIL and one CT module containing constructed response items) were not revised and were identical to the content presented in ICILS 2018. Further details of the development and revision of the ICILS 2023 main survey scoring guide for open-ended response items are provided in [Chapter 2](#).



The main survey scorer training employed a sample set of student responses collected during the field trial in English-speaking ICILS 2023 countries. The example responses used during scorer training were a mixture of those that clearly represented the scoring categories and those that were relatively difficult to score because they were partially ambiguous, unusually expressed, or on the “borderlines” of scoring categories. The scores that national center staff gave to these example responses were shared with the group, with discussion focusing on discrepancies in particular. The scoring guides and practice responses were refined following the scoring training to clarify areas of uncertainty identified during the scorer training.

Once training had been completed, the ISC provided national centers with a final set of scored sample responses as well as the final version of the scoring guide. The scored sample responses were accessible electronically through the web-based Assessment Master Marker module and were available only in English. National centers used this information, as they saw fit, to train their scoring staff on how to apply the scoring guides to the constructed-response items and large tasks. In some cases, national centers created their own sets of example responses from the student responses collected in their country.

To prepare for this task, the ISC provided national centers not only with suggestions on how to organize staff, but also with materials, procedures, and details on the scoring process. The ISC encouraged the national centers to hire scorers who were attentive to detail, familiar with education, and who, to the greatest extent possible, had a background in CIL. The ISC also provided guidelines on how to train scorers to accurately and reliably score the items and tasks.

Documenting scoring reliability

Documenting the reliability of the scoring process within countries was a highly important aspect of monitoring and maintaining the quality of the ICILS 2023 scored data. Scoring reliability within each country required two different scorers to independently score a random sample of approximately 200 responses for each constructed-response item and each large task.

The selection of responses to be double-scored and the allocation of these responses to scorers were random and managed by the web-based scoring software. The software was set up to ensure that a random selection of approximately 250 responses of all responses were double-scored, and that scoring could begin before all student responses had been uploaded to the system (thus allowing for late returns of data from some schools). The software set-up also allowed these tasks to be accomplished without compromising the selection probability of each piece of work for double scoring.

The degree of agreement between the scores, as assigned by the two scorers, provided a measure of the reliability of the scoring process. The web-based scoring system was able to provide real-time inter-rater reliability reports to scoring leaders who were encouraged (but not required) to use this information to help them monitor the quality of the scoring. Scoring leaders could, for example, use the information to monitor the agreement of each scorer with their colleagues (and identify scorers whose agreement was low relative to others), or identify items or tasks with relatively low inter-rater reliability that might need to be rescored or to have scorers provided with some additional training to improve the quality of their scoring.

Items with relatively low inter-rater reliability within a given country were not used in the estimation of student achievement for that country. [Chapter 11](#) outlines the adjudication process relating to inter-rater reliability.

Online data collection for survey activities questionnaires

In order to collect feedback about survey operations from NRCs, the international project team set up a survey activities questionnaire online. The questionnaire was prepared and administered using the IEA OSS. The survey activities questionnaire, unlike the other ICILS 2023 questionnaires, did not require national adaptations and was completed in English.

The purpose of the survey activities questionnaire was to gather opinions and information about the strengths and weaknesses of the ICILS 2023 assessment materials (e.g., test instruments, manuals, scoring guides, and software) as well as countries' experiences with the ICILS 2023 survey operations procedures. NRCs were asked to complete these questionnaires with the assistance of their data managers and the rest of the national center staff. The information was used to evaluate survey operations. It is also being used to improve the quality of survey activities and materials used in future ICILS cycles. Results of the survey activities questionnaire are reported in [Chapter 9](#).

IEA sent the NRCs individual login information and internet links for accessing the online questionnaires. Before submitting the responses to IEA, NRCs could go back and change their answers if necessary.

8.5 Field trial procedures

The ICILS 2023 field trial was a smaller administration of the ICILS 2023 assessment; on average, approximately 1,000 students were tested in each participating country.

The international field trial was conducted from May to June 2022.

The field trial was crucial to the development of the ICILS 2023 assessment instruments and also served the purpose of testing the ICILS 2023 survey operations procedures in order to avoid any possible problems during the ICILS 2023 data collection.

The operational resources and procedures described in this chapter were used during the field trial under conditions approximating, as closely as possible, those of the main survey data collection. This process also allowed the NRCs and their staff to acquaint themselves with the activities, refine their national operations, and provide feedback that could be used to improve the data collection procedures. As a result of the field trial, several modifications were made to survey operations, including enhanced procedures for collecting detailed information on the types of accommodations used for student instruments, where applicable. These adjustments supported the effective implementation of ICILS 2023.

Chapter 9:

Quality assurance in the ICILS data collection

Kateřina Hartmanová, Lauren Musu, and Marta Moreno Hidalgo

Introduction

To allow for ICILS 2023 data of the highest standard and comparability across countries, a significant effort has been put into ensuring standardized materials and survey operations procedures. For this purpose, quality control and assurance procedures were an integral part of ICILS 2023 and encompassed internal mechanisms built into each stage of the ICILS 2023 implementation.

The aim of this chapter is to provide an overview and present results of the quality assurance activities that occurred during the data collection for ICILS 2023. There are two primary activities described in this chapter. The first activity, the quality control program, consisted of two independent components: An international quality control program, managed by IEA, and a national quality control program, managed by each national center, based on the international guidelines and manuals provided.

The second activity that collected information on the overall quality of the ICILS 2023 administration was the Survey Activities Questionnaire (SAQ). Through this questionnaire, National Research Coordinators (NRCs) provided information about the implementation of the ICILS procedures, and the extent to which procedures were followed according to international standards.

More details on the quality assurance procedures employed during other phases of the ICILS 2023 study cycle—such as instrument development and preparation, scaling, and data analysis—can be found in the other chapters within this report.

9.1 International quality control program

This section provides an overview of the international quality control program and the data collected for the program.

The international quality control program was implemented by independent international quality control monitors (IQCMs) appointed and trained by IEA for each participating country. The major task of the IQCMs was to conduct on-site visits during the main survey data collection. In each country, the IQCM and the IQCM's assistant(s) visited a sample of testing sessions at participating schools. A total of 15 school visits were required unless specified otherwise. Other essential responsibilities of the IQCMs included visiting the National Study Center (NSC) for their country, meeting with the National Research Coordinator (NRC), interviewing the school coordinator and test administrator in each visited school, and reviewing selected sections of the *School Coordinator* and *Test Administrator Manuals*. Together, these tasks served to document data collection activities and determine whether the ICILS assessment was administered in compliance with the standardized procedures in their respective countries.

It is important to note that IQCMs visited only a sample of the schools that administered the ICILS survey in each country. In addition, IQCMs reported on any issues that may have been observed, even if only one student was affected. Results presented in this chapter should therefore be considered with this information in mind.

Selecting and training international quality control monitors (IQCMs)

The ICILS 2023 IQCMs were independent individuals experienced within school environments and (ideally) the information and communication technology (ICT) context, who resided in the participating country and had no affiliation with the NSCs. If no previous IQCM was available to serve again for



the country, IEA asked NRCs to nominate two candidates to serve as IQCMs for ICILS 2023 in their respective countries, which included sending the candidates' CVs and a short explanation why they thought the nominees would be suitable for the position at IEA. Potential candidates needed to be fluent in English as well as the administered language(s) in the target country. IQCMs were also required to not have any professional or personal affiliation with the NSC. Potential candidates were, for instance, school inspectors, relevant ministry officials, retired school teachers, or graduate students. Out of the IQCMs ultimately selected for ICILS 2023, nearly 60 percent previously served as an IQCM for another IEA study and the rest were newly nominated by NRCs.

Prior to beginning their assignments, the IQCMs were required to attend an online training session conducted by IEA. The online training session was held over two days in early February 2023. During the training, IQCMs were introduced to the ICILS operation procedures, study instruments, and tasks associated with their role. IQCMs were also provided with a manual detailing their responsibilities and the necessary materials for completing their tasks. After the training seminar and during the IQCMs' fieldwork, IEA remained in close contact with them and provided further assistance and instructions if required, especially in case of unforeseen circumstances.

For ICILS 2023, a total of 34 IQCMs were trained across 34 countries and one benchmarking participant.¹⁹ The benchmarking entity shared the same IQCM with their respective country. In turn, the IQCMs trained 90 assistants. Altogether, IQCMs and their assistants observed 509 administration sessions.

IQCM's tasks and responsibilities

The IQCMs were required to:

- Work with the NRC to select schools to visit to observe testing sessions and interview the school coordinator and test administrator
- Collect selected materials being used in their country to conduct the testing, including the *School Coordinator* and *Test Administrator Manuals*
- Compare selected sections of the national version(s) of the *School Coordinator* and *Test Administrator Manuals* against the international versions and document any substantial differences
- Submit electronic versions of the national version(s) of the above manuals and the results of the comparison to IEA
- Observe the ICILS 2023 testing sessions
- Verify the completeness and accuracy of the lists of participating classes, teachers, and students for each visited school
- Interview the school coordinator about ICILS 2023 activities leading up to and including the test administration
- Interview the test administrator about the actual administration of ICILS 2023
- Record the observations and the interview responses on the Classroom Observation Records (CORs) and submit these to IEA

Observing the ICILS administration sessions

IQCMs were required to document their observations of testing sessions using standardized CORs. The observation records were organized into four sections:

¹⁹ Please note that throughout this chapter we will use the term 35 countries when referring to the 34 countries and one benchmark participant.

- Section A: Preliminary activities;
- Section B: Administration process;
- Section C: Summary observations; and
- Section D: Interview with the school coordinator and the test administrator.

In Section A of the COR, IQCMs registered their observations of the condition of the assessment materials and testing equipment. Section B addressed different aspects related to the administration of the test and questionnaire, including how accurately the test administrator followed the prescribed administration procedures and read the administration scripts. Section C collected IQCMs' general impressions of the ICILS 2023 assessment and Section D was dedicated to the interview(s) with the school coordinator and test administrator, to get their evaluation of the ICILS 2023 assessment, gather suggestions for improvement, and obtain additional background information.

While IQCMs could decide for themselves how they recorded their findings during the school visits, they were required to enter the results of their observations and their notes in an online version of the COR after every school visit. The online version of the COR was accessible through an online platform and all information needed to be entered in English. IEA received documentation of the international quality control program from all 35 countries in ICILS 2023. The results of the IQCM observations are reported in the following sections. The tables and bulleted text throughout this chapter presents the IQCM responses to the questions in the COR.

9.2 Findings from the ICILS 2023 administration observations

Preliminary activities and condition of materials for the assessment

In the first section of the COR, IQCMs noted their observations of the condition of the assessment materials, the testing equipment, and the test administrator's level of technical preparation.

In line with Table 9.1, the survey materials were stored safely and securely in all cases where applicable (i.e., where physical materials such as USB sticks were used). Besides the test administrator, school coordinators and IT support staff often had access to the assessment room prior to the testing session. In addition, other individuals such as (vice) principals, the NRC, teachers of the class, handymen, or education department representatives could also access the room in some cases.



Table 9.1: Quality of ICILS 2023 administration preparation - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)	Not Applicable (%)
Were the assessment materials (e.g., USB sticks) safely and securely stored?	71	0	0	29
Besides the test administrator, did any of the following people have access to the assessment room prior to the testing session?				
<i>School coordinator</i>	84	16	0	-
<i>IT support person</i>	64	35	1	-
<i>Other</i>	28	72	1	-
What kind of devices are used for the assessment?				
<i>Desktop computers</i>	58	42	0	-
<i>Portable computers (e.g., laptops) connected to a power plug</i>	49	50	1	-
<i>Portable computers (e.g., laptops) not connected to a power plug</i>	22	77	1	-
Did the test administrator set up all workstations in a way that the welcome screen is displayed on each of them prior to the students arrival?	82	17	0	-
Does the test administrator use a device to keep track of time while the testing session is in progress?	98	2	0	-

In terms of equipment, desktop computers were most commonly used for the assessment (58%), followed by portable computers connected to a power plug (49%), and portable computers not connected to a power plug (22%). A combination of equipment types was used in some administrations. Despite the fact that some computers were not connected to a power plug, no power failure issues were reported.

The workstations were set up by the test administrator in a way that the welcome screen was displayed prior to the students' arrival in 82 percent of the observed sessions. In the rest, the computers either had yet to be turned on when the students arrived (especially when personal laptops were used for the testing) or were set up by a designated IT support staff rather than the test administrator themselves. Student IDs were displayed on the login screens in 90 percent of the cases, with student names used in the remaining 10 percent.

Each testing session started with the administration of computer and information literacy (CIL) modules, followed by the administration of the student questionnaire. Countries participating in the computational thinking (CT) module implemented this after the student questionnaire administration. For the purpose of this chapter, IQCMs' observations from the CT modules administration are reported last, as not all countries administered this module for ICILS 2023.

The ICILS student assessment modules were self-timed (30 minutes each for CIL/25 minutes each for CT), with recommended breaks of maximum 5 minutes between each CIL module and the student questionnaire (administration of which was estimated at approximately 25 minutes). Typically, a longer break of 15–45 minutes followed before the administration of the CT modules (where applicable).

While the testing session was in progress, the test administrator used a device to keep track of time in

the majority of the observed sessions (98%). In the remaining two percent of the observed sessions, the timer on the screens or a clock in the classroom were used, in other instances the time keeping task was delegated to somebody else (e.g., an assistant or the IQCM themselves). The modules were also self-timed, so that students could track their own progress.

Administration of the CIL module

Table 9.2 shows that IQCMs reported that the assessments were conducted in accordance with the international procedures. In very rare instances, the IQCMs observed and carefully documented significant deviations from the administration procedures which will be noted throughout this section.

In the majority (94%) of observed sessions, the test administrator ensured that each student sat in front of the computer specially prepared for him or her, sometimes with the help of the class teacher, by providing and checking the names/IDs written down on or in front of each computer or sending them directly to their allocated seats (in case they were familiar with the students). Alternatively, ID-s/login details were handed out after the students took their seats. Each student's participation status was recorded by the test administrator on the student tracking form in 97 percent of the observed administration sessions.

Table 9.2: CIL Test Administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)
Is the test administrator ensuring that each student is sitting in front of the computer specially prepared for him or her?	94	5	0
Does the test administrator record each student's participation status on the student tracking form?	97	2	1
Does the test administrator follow the test administration script in the ICILS 2023 <i>Test Administrator Manual</i> ?	72	28	0
Did any students finish the CIL test early (before the time allowed was up)?	93	7	0
Did students experience any other problems with the CIL testing?	36	64	0

One of the most important methods of standardizing the assessment administration was to have all test administrators follow the script in the *Test Administrator Manual*. IQCMs reported that in about three-quarters (72%) of the observations, the test administrators exactly followed the script when preparing the students, distributing the materials, and giving directions and examples. In the cases where the test administrators deviated from the script (28% – 141 cases total), most modifications were reported as “minor” (105 cases) and characterized most frequently as omissions (65 cases), followed by revisions (63 cases) and additions (61 cases). Note that these categories were not mutually exclusive and selecting multiple of them to characterize the changes (additions, revisions, omissions) was possible.

When IQCMs were asked to report on how many students did not have enough time to complete the CIL test module for each of the observer sessions:

- “None” of them did not have sufficient time to complete the CIL test modules in 54 percent of the observed sessions.
- “Some” of them did not have sufficient time to complete the CIL test modules in 43 percent of the observed sessions.
- “Many” or “all or almost all” of the students did not have sufficient time to complete the CIL test modules in two percent and one percent of the observed sessions.

Similarly, there were at least some students who finished the CIL test early (before the allotted time was up) in most (93%) of the observed testing sessions. These students were instructed to not disturb others, remain seated and keep quiet, and read or engage in other silent activities.

As for the CIL test modules content, IQCMs reported they observed that some (43%), many (2%), or all or almost all (1%) of the students found it difficult or confusing to answer the questions in the observed testing sessions, while none of the students appeared to have difficulty in the remaining 54 percent. Other than that, IQCMs stated that students may have experienced minor issues in about 36 percent of the observed sessions. These consisted mostly of technical issues such as weak or slow network connection, network breakdowns causing login problems, or interruptions during the test and computer or device malfunctions (e.g., USB issues, devices freezing or crashing, keyboard or mouse not working properly). In the majority of observed sessions, this impacted one or two students and the issues could be resolved during the administration such that the student could continue with the assessment. IQCMs also reported that some students experienced problems understanding and navigating the test according to the instructions (e.g., task skipping, understanding specific tasks), for a few of the students this was related to the language of the assessment not being their native language. Additionally, in some instances, students displayed a lack of motivation or did not seem to take the test seriously.

Table 9.3: Break after CIL test administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer/Not Applicable (%)
Is there a break between the end of the testing session and the start of the student questionnaire administration?	81	19	0
Are the computers and USB sticks (if applicable) secured after the CIL test administration during the break?	76	7	17

In accordance with the ICILS procedures, test administrators were asked to secure the test devices and USB sticks (if applicable) at the end of the testing session (see Table 9.3). The IQCMs reported that this occurred in 76 percent of the observed CIL administration sessions. In the remaining sessions, there was a general emphasis across the IQCM responses that at no point were the computers or testing materials left unattended, as there was always some form of supervision in place—the test administrators, students, or any additional personnel often remained in the room during breaks, meaning there was no perceived need to secure the materials. In addition, many responses indicated that breaks were either very short or non-existent. In cases where breaks were given (81% of the observed cases), they were often too brief to require securing the materials. Several responses also noted that the test was conducted online, which reduced concerns about the physical security of materials such as USB sticks or paper tests. The IQCM responses collectively highlighted a consistent approach to ensuring test security through supervision and procedural measures, rather than securing materials in the traditional sense.

Administration of the student questionnaire

Table 9.4 shows selected questions from IQCMs' observations of the student questionnaire (StQ) administration. IQCMs reported that the test administrator recorded each student's participation status for the StQ on the student tracking form in 95 percent of the observed sessions. For the remaining five percent, issues with the participation status were resolved at a later stage by IEA together with the NRC.

As for following the StQ administration script in the *Test Administrator Manual*, IQCMs reported that in about three-quarters of the observations (78%), the test administrators followed the script exactly. In the cases where the test administrators deviated from the script (21%—109 cases total), most modifi-

cations were also here reported as “minor” (75 cases) and characterized most frequently as omissions (61 cases), followed by revisions (46 cases) and additions (32 cases). Note that these categories were not mutually exclusive and selecting multiple of them to characterize the changes (additions, revisions, omissions) was possible. Most commonly, these modifications were described as either simplification of the language and instructions used to ensure students understood them, omissions of repeated content and irrelevant details, using own words to rephrase the scripts to avoid repetition, clarifying potentially confusing instructions, offering extra support to students in case of need, and adding specific technical instructions or adaptations of the script to fit the local context.

Table 9.4: Student questionnaire administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)
Does the test administrator record each student's participation status for the student questionnaire on the student tracking form?	95	4	1
Does the test administrator follow the questionnaire administration script in the ICILS 2023 <i>Test Administrator Manual</i> ?	78	22	1
Do the students ask for additional time to complete the questionnaire?	21	79	1

In about 21 percent of the observed sessions, students asked for additional time (beyond the estimated 25 minutes) to complete the questionnaire. There was a break given between the end of the questionnaire administration and the start of the CT test administration in 67 percent of the observed sessions (note this was only applicable where the CT module was administered—i.e., in 23 out of the 35 participating countries). The most common reasons reported for not having a break were the students' preference to continue without one, the fact that many students had finished the questionnaire earlier making a break unnecessary, having a longer break earlier, or practical constraints (such as issues with scheduling and technology).

Table 9.5: Break after student questionnaire administration - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)
Is there a break between the end of the questionnaire administration and the start of the CT test administration?	67	10	23
Are the computers and USB sticks (if applicable) secured at the end of the questionnaire administration?	91	8	1

As presented in table 9.5, the computers and USB sticks used were secured at the end of the questionnaire administration in 91 percent of the observed sessions (where applicable), as per IQCM reports. In the sessions where this procedure was reportedly not followed, the IQCM responses commonly pointed out that test administrators or other personnel often remained in the classroom to monitor the materials and students were instructed not to use the computers (which were often secured with login codes), so there was no perceived need to further secure the materials. In some cases, students and test administrators both left the classroom, but the doors were typically closed, locked, or monitored. In the sessions where no break was given, the CT module administration followed immediately, avoiding the need to secure the materials between one session and the other.

Administration of the CT module

Table 9.6 illustrates IQCMs' observations of the CT module administration. Only the 355 observed sessions that included CT administration were included in the analysis. IQCMs reported that the test administrator ensured that all students were sitting at the correct computers in 97 percent of the observed sessions. In the two percent of the observed sessions where this procedure was reportedly not followed, the most commonly given reason was that this was not necessary as students remained seated during the break or there was no break altogether. In a few other cases, the test administrator trusted the students to come back to their assigned seats.

According to IQCMs' reports, the test administrators followed the CT administration script exactly in 82 percent of the observed sessions. In the cases where the test administrators deviated from the script (17%—60 cases total), most modifications were reported as "minor" (42 cases) and characterized most frequently as omissions (38 cases), followed by revisions (28 cases) and additions (10 cases). Note that these categories were not mutually exclusive and selecting multiple of them to characterize the changes (additions, revisions, omissions) was possible. Most frequently, the changes mentioned in IQCM reports were adaptations or simplifications made by the test administrator to fit the target group and classroom situation, omissions of details about the different modules, types of questions or procedural specifics like the timer, and rephrasing and summarizing of the instructions using the test administrator's own words.

Table 9.6: CT module administration - Across 355 observed ICILS administration sessions where CT was administered (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)
Does the test administrator ensure that all students are sitting at the correct computers?	97	2	1
Does the test administrator follow the CT test administration script in the ICILS 2023 <i>Test Administrator Manual</i> ?	82	17	1
Did any students finish the CT test early (before the time allowed was up)?	90	9	1
During the CT test administration, did students experience any other problems with the testing?	21	77	1

The IQCMs reported that all students had time to complete the CT test modules in nearly two-thirds (59%) of observed testing sessions. In the rest of the observed testing sessions, in less than one percent of the sessions all or almost all students did not have enough time to complete the CT test modules and some of them did not have enough time in 38 percent of the sessions. There were at least some students who finished the CT test early (before the allocated time was up) in most (90%) of the observed testing sessions. These students were instructed to not disturb others, remain seated and keep quiet, read or engage in other silent activities. In few cases, students were also allowed to leave the room or return to their regular classes after finishing early.

As for the CT test modules content, IQCMs reported that some of the students found it difficult or confusing to answer the questions in 47 percent of the observed testing sessions, many of them in six percent of the sessions, all or almost all in one percent of the observed sessions, and none in 44 percent of the sessions. Other than that, IQCMs reported that students experienced minor problems with the testing in about 21 percent of the observed sessions. These problems usually affected only one to two students per observed session and consisted mostly of problems with login, technical issues with the hardware, software or internet connection, problems with concentration or motivation (sometimes also exacerbated by a noisy environment created by other students who finished early) and problems with understanding specific task instructions.

Summary observations of the ICILS 2023 administration

Section C of the COR included questions about IQCMs general impressions of how overall the administration was conducted and how well the test administrator monitored students (Table 9.7), the students' cooperation (Table 9.8), and any special circumstances or technical issues (Tables 9.9 and 9.10) that arose during the session. The below tables report the IQCM responses on these.

Administration and test administrator monitoring

IQCMs perceived that most test administrators (96%) were familiar with the test administration script. In addition, a majority of the observed administration sessions took place under favorable room conditions that were suitable for students to work without distraction (91%), as indicated by the IQCMs. The most commonly reported inconveniences were noise distractions from outside the classroom, temperature issues (i.e., the classroom being either too hot or too cold), poor lighting or ventilation, at times lack of space between the individual desks (affecting 11% of the observed sessions, as reported in Table 9.7) or disruptions by the school personnel entering the room or talking among themselves. The test administrator walked around the room to be sure students were completing the test or the questionnaires and/or behaving properly in nearly all (98%) of the observed sessions. In the rest of the sessions (2%), the test administrator sat at their desk and monitored the students from there. According to the IQCMs' opinions, the test administrator addressed students' questions appropriately in 95 percent of the observed sessions. In the remaining cases, the test administrator struggled to answer the students' questions, gave answers that were too specific, or simply referred students to the task instructions.

Table 9.7: General IQCM impressions and observations - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer/ Not Applicable (%)
In your opinion, did the test administrator familiarize himself or herself with the scripts prior to the administration of the assessment instruments?	96	4	0
In your opinion, were the conditions in the testing room suitable (lighting, temperature, noise, etc.) for the students to work without distractions?	91	9	1
Did the seating arrangement provide adequate space for students to work and not be distracted by each other?	89	11	0
During the administration sessions, did the test administrator walk around the room to be sure students were completing the test or the questionnaires and/or behaving properly?	98	2	0
In your opinion, did the test administrator address students' questions appropriately?	95	2	3

Students' cooperation

Student refusal to take the assessment was rare and only occurred in six percent of observed assessment sessions (see Table 9.8), most often (in 3% of the observed sessions) only a single student within the classroom refused. There were a few problems observed in relation to test administrators accurately following the instructions for excusing the student who refused to take the test, which consisted either of the student not being responsive or the test administrator not making a record of this anywhere.

There were no late students in the majority (86%) of observed sessions. Out of the students who did arrive late (14% of observed sessions), most were still admitted into the room (10%), either before (6%) or after (4%) the beginning of the administration. Across all the observed sessions, there were 19



reported cases (4%) of late students not admitted to the room. In 17 percent of the observed administration sessions, students left the room for an “emergency” during the test administration (examples of “emergencies” included students needing to use the toilet, or feeling unwell at the time of the testing). In 83 percent of such cases, IQCMs reported that the test administrator addressed the situation appropriately. In the 15 percent of the cases where IQCMs did not perceive this was handled appropriately, the IQCMs reported that the test administrator did not take any action or that the computer was not secured during this time. However, IQCMs often clarified that it did not seem necessary to secure the computer.

Table 9.8: Student cooperation - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)
Did any students refuse to take the test either prior to the testing or during the testing?	6	94	0
If a student refused, did the test administrator accurately follow the instructions for excusing the student?	74	19	7
Were any late students admitted to the room?	10	90	0
Did any students leave the room for an “emergency” during the testing session?	17	83	0
If students left the room for an “emergency,” did the test administrator address the situation appropriately?	83	15	2
Did students store away everything, including all electronic devices, such as mobile phones, portable computers, and photo or video cameras, for the duration of the assessment session?	89	11	0
Did you see any evidence of students attempting to cheat on the test (e.g., by copying from a neighbor)?	8	91	0

While in most observed testing sessions (11%) students complied with the instructions to store everything away, in a few cases (7 reported cases—1% of observed sessions) this instruction was never given and students kept their cell phones on them. In some cases, students were allowed to use their phones or other devices after they had finished the test to keep them from disturbing others. In addition, the IQCMs noted that test administrators and teachers faced difficulties in consistently enforcing these rules. Repeated interventions by the test administrators were documented in a few cases, with some teachers resorting to a proactive approach of collecting the phones from students. In addition to cell phones, some students also reportedly had tablets, e-readers, or laptops on their desks.

In eight percent of the observed sessions, IQCMs reported seeing evidence of students attempting to cheat, such as glancing at fellow students’ screens, discussing among themselves or using their phones when the test administrator wasn’t looking. IQCMs reported that test administrators intervened in appropriate ways whenever they noticed this. However, the IQCMs also pointed out that the assessment was different for individual students, which made it unlikely for them to effectively cheat.

Overall, IQCMs described the students as:

- “Extremely orderly and cooperative” in 60 percent of the observed sessions
- “Moderately orderly and cooperative” in 31 percent of the observed sessions
- “Somewhat orderly and cooperative” in eight percent of the observed sessions
- And “Disorderly and hardly cooperative” at all in the remaining two percent of the observed sessions

The additional comments on student behavior during the testing reflect the above data and reveal a mix of positive and less positive observations. While many students were described as extremely orderly, cooperative, and well-behaved, other students were described as restless, unmotivated, and slightly disruptive. In many cases, the combination of fatigue, technical difficulties, and early finishers led to some level of disorder.

Special circumstances

There were students using special accommodations (additional support) to complete the instruments in 12 percent of the observed sessions (Table 9.9). Several students were reportedly granted additional time, with specifics including 10 extra minutes (7 cases), 20 extra minutes (1 case), and multiple unspecified instances. For other students, the timer was disabled altogether, which in a few cases turned out to be unnecessary as they completed the test before the allotted time ran out. Multiple students completed their tests in a separate room, often with an additional test administrator or teacher present to help, while others received help from a teacher or assistant in reading instructions and/or writing their responses in the classroom. In addition, a tablet was used to translate test content into another language for some students, and special information or translation assistance was provided for language reasons in a few cases. There were also reported cases of students with a social-emotional support assistant in the room to manage behavior or students using headphones during the test.

Table 9.9: Special accommodations - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)
Did any students use accommodations (additional support) to complete the instruments?	12	88	0
Were the accommodations provided to the qualified students consistent with the agreed accommodations listed on the student tracking form?	73	22	5

Where accommodations were provided, these were reported to be consistent with the agreed accommodations listed on the student tracking form in 73 percent of the observed cases. In the remaining 22 percent of reported cases, the inconsistencies mentioned were either cases where students were listed to require special accommodations but were effectively not using them, or not listed but using accommodations regardless. The need for accommodations was often only identified on the day of the assessment, contrary to the student tracking form which listed that no accommodation was necessary. In one case, it was reported that a test administrator attempted to provide the accommodation (disabling the timer), but was not successful.

Technical challenges

There were technical problems reported with using computers in 27 percent of the observed sessions, using USB sticks in six percent of the observed sessions, logging into the assessment player in 12 percent of the observed sessions, timer function in two percent of the observed sessions and other malfunctions when using the assessment player in 23 percent of the observed sessions (see Table 9.10). Most commonly mentioned were login and connectivity issues, applications or computers freezing or crashing, language or display issues, hardware malfunctions, and timer or other assessment player issues. Within the observed sessions where technical problems occurred (47% of the observed sessions), the IQCMs reported that the test administrator solved those problems effectively in 85 percent of the relevant cases. However, in the remaining 15 percent of the relevant cases (7% of all observed sessions), test administrators were not always able to solve the problems effectively or the solution took longer than expected. There were cases (although these were the exceptions) where despite repeated attempts and contacting IT support the issue could not be resolved. A member of IT staff was available to provide support throughout the ICILS administration process in 71 percent of

the observed cases.

As in other instances, IQCMs also reported that technical issues arose only with one to two students per observed session. In addition to that, IQCMs also mentioned that overall when these issues arose, test administrators and IT support, were able to quickly address these, minimizing the overall impact on the testing.

Table 9.10: Technical issues and support - Across all 509 observed ICILS administration sessions (percent of IQCM responses, rounded to a whole number)

	Yes (%)	No (%)	No Answer (%)	Not Applicable (%)
Were there any technical problems with the procedures or devices below?				
<i>Using computers</i>	27	72	0	1
<i>Using USB sticks</i>	6	33	0	61
<i>Logging into the assessment player</i>	12	83	1	4
<i>Timer function</i>	2	95	1	2
<i>Other malfunctions when using the assessment player</i>	23	72	1	5
If there were any technical problems, did the test administrator solve them effectively?	40	7	0	53
Was an IT staff member available to provide support throughout the ICILS administration process?	71	29	0	-
Did the test administrator submit the data from each USB stick used for the testing session directly after the testing session?	21	6	0	73

In over three quarters (78%) of the relevant observed sessions, the IQCMs reported that the test administrator submitted the data from each USB stick used for the testing session directly after the testing session. In the 21 percent of the relevant sessions where this was not done, the IQCMs described that the test administrator either planned to do this later in the day (sometimes after another testing session) or that it was the NSC who would be submitting the data, and the USB sticks were sent to them instead.

General impressions

When IQCMs were asked to report on the overall quality of the ICILS administration:

- Forty-five percent of the IQCMs described the quality as “Excellent”
- Thirty-five percent of the IQCMs described the session as “Very Good”
- Thirteen percent of the IQCMs described the session as “Good”
- Only four percent and three percent of the IQCMs described the session as “Fair” or “Poor”

The additional IQCM verbal feedback on the ICILS 2023 testing sessions highlighted that the administration process was generally deemed well-organized, with many praising the professionalism and preparedness of the test administrators and school coordinators. Students were described as mostly cooperative and engaged, although there were instances where the students displayed challenging behavior or rushed through the tests. In this line, IQCMs also reported that some students showed some signs of fatigue due to the length of the testing sessions. In some other instances, some technical

issues slightly impacted the smooth running of some observed sessions.

Table 9.11: Administering the main survey - Accommodations (number of NRC responses, max 35)

	Not difficult at all	A little difficult	Somewhat difficult	Very difficult
How difficult was it to manage the different steps in the process of handling accommodations for qualified students? (if applicable)				
Having the school coordinators provide details of students qualified to receive accommodations on the student listing form.	11	5	4	1
Confirming with school coordinators that nominated students were qualified to make use of accommodations.	1	6	3	1
Using the accommodation codes on the student listing and tracking forms.	11	7	4	0
Making arrangements with schools and test administrators regarding the implementation of listed accommodations on the days of testing.	10	4	6	1
Making use of the function to disable the timer (where appropriate) on the days of testing.	13	2	6	0
Keeping records of the additional time provided to students who qualified for and made use of this accommodation.	10	3	2	6

Interview with the school coordinator and test administrator

The IQCMs conducted interviews with the school coordinators and test administrators according to the guidelines included in Section D of the COR. The purpose of this activity was to solicit an evaluation about the ICILS administration from the responsible persons, to collect suggestions for improvement, and to obtain additional background information on survey-related activities such as:

- Shipment of assessment materials
- Arrangements for ICILS administration
- Responsiveness of the national center to queries
- Organization of classes in the school (to validate within-school sampling procedure)

School coordinators

When school coordinators were asked about their position at that school/school district, it was reported that:

- Thirty-two percent worked as assistants or deputies to the principal/school head.
- Twenty-one percent worked as a teacher from that school, (although not as a teacher of the class assessed).
- Twenty percent acted as principals or school heads.
- Eleven percent worked as the teachers of the assessed class.
- Fifteen percent held other positions at the school or school district.



- One percent acted as counselors for that particular school.

The school coordinators rated the attitude of the school staff members towards ICILS 2023 as positive in 62 percent of cases, neutral in 31 percent, and negative in the remaining seven percent of the cases. The related comments showed that the majority of staff members were positive and supportive of the assessment, appreciated its significance and were willing to contribute in most cases. Especially schools with a focus on digital literacy viewed the assessment as relevant and were curious about the results, and staff members valued the professionalism and the experience gained from participating in an international assessment. In addition, some staff members were already familiar with such tests, finding no major issues with participation. Nonetheless, a portion of staff often viewed the assessment as an additional burden due to its timing and the preexisting workload. The timing of the assessment was commonly criticized as it coincided with national exams and the end of the school year, leading to some stress and inconvenience. Furthermore, some teachers felt that the assessment affected regular teaching schedules and added pressure during busy periods. Some other points of concern were the lack of (immediate) feedback from the assessment, logistical challenges in organizing the test, or doubts about the necessity and benefits of the assessment. Overall, while there was some resistance and fatigue due to the additional workload, the prevailing attitude reported was cooperative with an understanding and acknowledgment of the test's importance.

As for school coordinator preparations prior to the assessment day, there were a small number of cases where some components from the shipments of test materials were not received by the school coordinator, as Table 9.12 shows. In many instances, USB sticks were brought by test administrators on test day, and schools often did not receive return envelopes or boxes. Common explanations for this were that the process was managed electronically, reducing the need for physical materials, or that the test administrators (rather than school coordinators) were responsible for receiving and managing test materials, including manuals and USB sticks. In a few cases, school coordinators received the shipment from the NSC but did not open or inspect them due to time constraints or perceived irrelevance, while other school coordinators were still waiting to receive (some of) the materials at the time of reporting. Many responses also indicated that external organizations, such as the NSC or contracted companies, handled significant parts of the process, including material distribution and collection. In one country, the *School Coordinator Manual* was not used as such, instead, a curated website designed by the national center was shared with school coordinators via email.

School coordinators reported that the NSC was responsive to questions and concerns in 89 percent of the cases. In most other cases, the school coordinators mentioned delays and fluctuations in communication, depending on the time period. In line with technical preparations and procedures, 79 percent of the school coordinators reported they had run the ICILS Compatibility Check Program on at least some of the computers prior to the assessment. In the cases this was not done (21%), the reasons provided included reliance on external support for this (from the test administrator, NSC, or IT specialists), lack of information or awareness of its necessity, logistical constraints (e.g., lack of time or access to the devices prior to the assessment), assumptions of pre-checked equipment, and either having conducted other (partial) checks or deeming the compatibility check unnecessary (e.g., due to the use of online testing platforms).



Table 9.12: Interview with the school coordinator - Preparations (percent of school coordinator responses)

	Yes (%)	No (%)	No Answer/Not Applicable (%)
Did you receive the correct shipment of the following items?			
<i>School Coordinator Manual</i>	92	3	0
<i>Test Administrator Manual</i>	73	12	16
<i>All USB sticks for sampled students</i>	28	6	66
<i>Teacher tracking form</i>	94	3	3
<i>Student tracking form</i>	96	2	3
<i>Test administration form</i>	77	8	14
<i>Envelopes or boxes addressed to the National Center for the purpose of returning the materials after the assessment</i>	52	13	35
Was the National Study Center responsive to your questions or concerns?	89	2	10
Did you run the ICILS Compatibility Check Program on the computers prior to the ICILS assessment?	79	21	0
Were you satisfied with the set up that you were able to arrange for the administration of the assessment and questionnaire?	95	5	0

Overall, a vast majority (95%) of school coordinators were satisfied with the setup they were able to arrange for the administration of the assessment and questionnaire. The reasons for dissatisfaction (5%) included technical issues (e.g., unreliable computers, software issues, confusion about compatibility checks), logistical challenges (e.g., changes in test dates, tight schedule, time and administrative constraints), coordination or communication issues, insufficient room preparation or control over students, and concerns about inclusivity.

As for incentives, teachers, ICT coordinators and/or the school principal were reported to receive incentives for participating in ICILS 2023 in 13 percent of the visited schools. The most common incentives for school staff included monetary compensation, gift cards, and snacks or meals during breaks. Certificates and small gifts were also frequently mentioned, alongside professional credits and benefits (e.g., free time, access to webinars). While many incentives were provided, there were several instances where promised incentives had not yet been received at the time of the interview.

When school coordinators were asked to report on some of the forms used during the test administration operations (see more information about this in [Chapter 8](#)), 88 percent of the coordinators provided the list of classes in the tested grade to the NSC and 97 percent of the IQCMs reported that the information provided by them matched the information provided on the listing form (see [Table 9.13](#)). The only deviation mentioned was a case of the student listing form not matching the identification of students excluded with Special Educational Needs on the registration form. To the best of school coordinators knowledge, there were no students in this grade level enrolled in more than one of the grade 8 classes that they listed in 99 percent of the cases. The reported exceptions were students enrolled in special programs to support them due to their academic difficulties. All students participating in the ICILS assessment were listed on the student listing form in 90 percent of the reported cases. The students not listed on the listing forms were predominantly due to absences (sickness, personal reasons), new arrivals or transfers, exclusions by parents or due to behavioral issues, and tracking form discrepancies.



Table 9.13: Interview with the school coordinator - Listing forms (percent of responses)

	Yes (%)	No (%)	No Answer (%)
Did you provide the list of classes in the tested grade to the national center?	88	12	0
(IQCM response) Does the information provided by the school coordinator match the information provided on the listing form?	97	3	0
To the best of your knowledge, are there any students in this grade level enrolled in more than one of the grade 8 classes that you listed?	1	99	1
Were all students participating in the ICILS assessment listed on the student listing form?	90	10	0
Did the teacher listing form include all eligible teachers for the target grade?	93	7	0

The teacher listing form included all eligible teachers for the target grade in 94 percent of the reported cases. The primary reasons for teachers not being listed on the forms included administrative oversights, misunderstandings, sampling processes, teacher availability, and lack of clarity or time to include all teachers. In a few cases, the school coordinator had not compiled the list yet or did not know why certain teachers were not included.

Only a bit over third (35%) of the school coordinators were reportedly involved in the selection and training of test administrators at their school (see Table 9.14). All (100%) of the relevant school coordinators reported this went well at their school and nearly all (98%) reported the *Test Administrator Manual* was useful for the purpose. Out of the relevant school coordinators, 84 percent were given time by the principal to do the training. Some school coordinators expressed a need for more efficient training provided by the NSC due to time constraints, especially for those with additional responsibilities. Both school coordinators and test administrators often had to self-train using the provided manuals. Eighty-six percent of school coordinators expressed that they felt the ICILS 2023 *School Coordinator Manual* worked well. Suggestions for improvement included more efficient organization of the manuals, better scripting, and more distinctive font usage to enhance readability. In addition, making the manual shorter, simpler, with clearer instructions would have been appreciated. Enhancing the training and communication, addressing specific instructional gaps, and tailoring content to country-specific needs were also highlighted as important improvements to the process. Overall, while there were some challenges related to time constraints, material clarity, and role overlap, school coordinators generally managed to prepare and select test administrators effectively through collaboration, self-training, and external support.

Table 9.14: Interview with the school coordinator - test administrator training (percent of school coordinator responses)

	Yes (%)	No (%)
Were you involved in the selection and training of test administrators at your school?	35	65
(If yes)		
Did the selection and training of test administrators go well?	100	0
Was the <i>Test Administrator Manual</i> useful for training purposes?	98	2
Were you given time by the principal to do the training?	84	16

As far as the school coordinators were aware, the respondents encountered problems filling in the ICT coordinator questionnaire in seven percent of the reported cases, the teacher questionnaire in 14

percent of the reported cases and school questionnaire in eight percent of the reported cases (see Table 9.15). The main issues included technical problems (e.g., issues with access or login, missing email invitations, submission errors), complex and lengthy content (e.g., confusing phrasing, too time-consuming), inadequate translations, and procedural misunderstandings (e.g., unclear instructions, lack of feedback). A call for a simpler and less time-consuming process for filling out questionnaires was evident and the ability to make remarks in an open question format was requested to better explain conditions and responses. In addition, the need for more specific and relevant questions tailored to local contexts and education systems was highlighted.

Table 9.15: Interview with the school coordinator - Final observations (percent of responses)

	Yes (%)	No (%)	No Answer (%)
As far as you are aware, did the respondents encounter any problems filling in the background questionnaires?			
ICT coordinator questionnaire	7	91	2
Teacher questionnaire	14	85	1
School questionnaire	8	91	1
Would you be willing to serve as a school coordinator again?	90	10	0

Altogether, and concerning the overall quality of the ICILS 23 administration:

- Seventy-six percent of the school coordinators reported that the administration in their school went “overall very well”
- Twenty-two percent of the school coordinators reported that the administration went “satisfactorily with few problems”
- Only two percent of the school coordinators indicated that the administration went “unsatisfactorily and with many problems”

In line with the coordinators’ positive impressions about the study administration, the vast majority (90%) reported that they would be willing to serve in this role again.

Additional feedback provided by the coordinators reflected a mix of positive and critical insights about the ICILS testing process. While many appreciated the opportunity for international experience and praised the organization and support from the NSC, there were also some worries about the time and effort required for preparation and administration, which at times impacted other school activities and caused some stress for both students and staff. Issues with the technology, the timing of the tests, and the complexity of questionnaires were common themes. School coordinators suggested the need for better timing in the administration of the study, ideally not overlapping with busy periods such as exams or holidays, the provision of more streamlined and accessible training and materials, improved technical support, and more flexibility to accommodate specific school contexts circumstances.

Test administrators

IQCMs also interviewed the test administrators—in 17 percent of the cases this was the same person as the school coordinator. About 22 percent of test administrators were teachers from a different class, nine percent teachers from the assessed class, four percent assistants or deputy principals/school heads, two percent principals/school heads, one percent school counselors and the rest (63%) fell into the “other” category—mostly external individuals not affiliated with the schools being assessed. They were hired by various external entities such as the NSC or consulting agencies. Several test administrators held positions as educational methodologists (or were associated with methodological centers), other educational professionals (e.g., psychologists, retired educators, researchers) or student assistants.

Regarding test administrator preparations prior to the assessment day, there was a small number of cases where some materials were not received by the test administrator, as shown in Table 9.16. Nearly 83 percent of test administrators felt the ICILS 2023 *Test Administrator Manual* worked well, while the remaining test administrators indicated the manual needs improvement. The suggested improvements included for the manual to be simplified and made more concise, using language that is easy for students to understand and eliminating redundant information. Test administrators also suggested structure improvements such as providing separate guides for different tests, boxing read-aloud instructions, highlighting key information or providing other visual aids. The test administrators further expressed a need for clear timing instructions and strategies for managing student behavior, especially for early finishers and those requiring extra time. Additionally, it was mentioned that tips on addressing technical issues would improve usability. Providing supplementary materials like PowerPoint instructions and an FAQ section was also suggested to further support test administrators.

As for the setup of the administration of the assessment and questionnaire, 93 percent of the test administrators who were different from the school coordinator expressed their satisfaction with this (77% of test administrators overall). The remaining seven percent pointed out that the preparation and setup phase faced challenges due to technical issues, inadequate IT support, and cumbersome logistics (e.g., managing insufficient school staff and the timing of the questionnaire), which often fell solely on the test administrators' shoulders. Additionally, environmental factors and room arrangements hindered effective student monitoring and contributed to a sub-optimal testing environment. Behavioral issues and lack of student motivation further complicated the testing process.

Table 9.16: Interview with the test administrator (percent of test administrator responses)

	Yes (%)	No (%)	No Answer/Not Applicable (%)
Did you receive the following materials?			
<i>Test Administrator Manual</i>	100	0	0
<i>Test administrator summary sheet</i>	92	7	0
<i>Test administration form</i>	98	1	0
<i>USB sticks</i>	42	2	57
<i>Student tracking form</i>	98	2	0
Were you satisfied with the set up for the administration of the assessment and questionnaire?	77	6	17
Did you list any problems encountered during the ICILS administration on the test administration form?	36	30	35
Would you be willing to serve as a test administrator again?	92	8	0

Over one-third (35%) of test administrators reported there were no problems encountered during the ICILS administration. Out of the remaining 65 percent of test administrators, 36 percent reported any problems encountered were listed on the test administrator forms while the rest (29%) did not list those. The issues described in test administrator comments included technical problems such as browser crashes, internet connectivity issues, computer malfunctions, software freezes, and USB stick failures. There were also disruptions caused by late arrivals, traffic delays, and unexpected entries by teachers and students. In addition, some students faced difficulties with logging in, completing tasks, and understanding instructions due to some discrepancies in the manual and the tutorial. Some behavioral issues also arose, such as students not taking the survey seriously, logging out prematurely, or not finishing modules. Few of the reported problems were specific to task functionality (e.g., non-working backspace keys and unresponsive maps). Despite these challenges, most issues were reportedly addressed on the spot by IT support or solved by restarting the systems.

Overall, 67 percent of the test administrators who were different from the school coordinator said

the assessment administration went very well with no problems, 28 percent said it went satisfactorily with few problems and only five percent said it went unsatisfactorily with many problems, reflecting the issues mentioned above. Despite the challenges, 92 percent of the test administrators expressed they would be willing to serve as a test administrator again. Open feedback from test administrators highlighted several areas for improvement, as well as some positive aspects of the process. Many test administrators noted that the test duration was too long and suggested that longer breaks between modules or splitting the test into two parts could enhance focus and performance. There were concerns about the complexity and length of the instructions and manuals, which could be streamlined for clarity. Technical issues with equipment and logistical challenges underscored the need for better technical support and preparation. Test administrators also expressed a wish for more comprehensive pre-test training and a demo version of the test to better prepare themselves for student questions. On a positive note, the cooperation from NRCs and the support from IEA were appreciated, and the standardized approach was seen as beneficial. Overall, while the process was generally well-organized, adjustments in timing, training, and technical support were suggested to improve the testing experience.

9.3 Findings from the national materials review

As part of their tasks, IQCMs reviewed selected sections of the national versions of the *School Coordinator* and *Test Administrator Manuals*, to ensure the content was comparable to the international versions.

For the *School Coordinator Manual*, the selected section was one outlining the exclusion and accommodation rules that school coordinators should follow.

Twenty-five out of the 35 participating countries made some adaptations or omissions to the reviewed text. A majority of these were very minor and/or necessary to reflect the respective country's context, most commonly affecting the passages and codes detailing the specific accommodations made for students requiring additional support, often reflecting the fact that a type of support/accommodation was not available in the given country due to lack of resources or other organizational limitations. In five countries, special accommodations for students requiring additional support were not available at all. In addition, some omissions were made related to some parts not being applicable, e.g., related to a specific mode of administration (i.e., USB) in a given country. Another common omission was related to a paragraph describing how to fill in the column of a student listing form related to the language of the test, which was optional for school coordinators to fill in. It is also interesting to note that in one country the *School Coordinator Manual* as such was not used and was replaced by another resource (a curated website) providing the same information in what was considered a more engaging and digestible format.

For the *Test Administrator Manual*, IQCMs were asked to review a section including the script describing the assessment procedures to students, particularly the test introduction. There were adaptations or omissions to the reviewed text reported in 21 of the 35 countries. Omissions of full sections were very rare, only reported in two countries and both related to the visual coding tutorial. Adaptations to the test administrator script were more common and mostly related to the technical/procedural instructions, such as "You must click on the arrow to move to the next task" or "You cannot return to previous tasks." In these cases, the text was often rephrased or reworded (while the overall meaning remained), for example, replacing "must" with "need to" or simply the "click on" instruction. As reported by one of the IQCMs, "The main difference is the choice of words used to convey the same idea." In addition, specific instructions from the tutorial messages (such as "Go back" or "Try again") have been correctly adjusted to reflect the adaptations/translations used in the national materials.

Some additional fillers or omissions for test instructions reported more than once were leaving out a mention about the displayed websites being safe for students to use (which was already implied by the previous sentences), leaving out information on the breaks between modules and what to do once a student completed the assignment, leaving out the reminder to complete the modules independently



and not disturbing others, naming few examples of tasks students might encounter (e.g., visual coding, programming), or adding some motivational sentences such as “Please try to do your best in the tests. Please try to think of working on them as a kind of international competition in which you and pupils from other schools represent [name of the country]” and advising the test administrator “One way to do this is to start with a short talk about the importance of the work they [the students] are about to undertake.” Finally, some adjustments were made in a few countries related to how the timing of the test and tutorial was described, while still preserving the important information (such as changing “no more than 30 mins” to “exactly 30 mins,” or not specifically mentioning the 5 minutes inbuilt into the timer for the tutorial but highlighting that this timer is only there as an indication in this case and that the time for the tutorial is in fact not limited).

Some additions were also made to the StQ instructions, such as reminders that students did not have to answer all the questions if they did not wish to, that the results would not affect students’ grades, that it is not possible to trace a response back to the individual students, or that students should not eat or drink at the computers. Differences were also reported in the specific terms used to refer to the “parents/guardians” in the instructions, presumably also in line with the country-specific adaptations made to the national instruments.

In general, any modifications reported by the IQCMs to the national versions of the *School Coordinator* and *Test Administrator Manual* were minor and did not compromise the quality of the collected data.

9.4 Findings from the survey activities questionnaire

As another valuable source of information on the overall quality of the ICILS administration, the SAQ gathered information from the NRCs about whether the implementation of the ICILS procedures accorded with the standards outlined in the Survey Operations Procedures (SOP) manuals. The SAQ also collected feedback on the strengths and weaknesses of the approaches and materials used. It covered various topics including:

- sampling
- contacting schools
- recruiting school coordinators
- translating and preparing the survey instruments
- administering the assessment
- implementing the national quality monitoring program
- scoring open-ended response items
- entering and coding occupation data
- submitting data
- time required for survey activities
- other experiences

One major purpose of the SAQ is to gather information to improve the quality of future IEA studies. Data were collected online from each NRC personally, with data manager and/or other national center staff assisting this process where necessary. The results presented in this section reflect the quality of the ICILS procedures and materials in all of the participating countries.

Sampling related procedures

The first part of the SAQ collected information regarding sampling related procedures. Table 9.17 provides the NRCs’ responses to questions about sample selection and other related matters. Out of the

35 respondents, 32 (all those who answered the question) indicated they felt well supported by the IEA sampling team. Concerning the Survey Operations Procedures Unit describing the sampling process, 32 NRCs felt it sufficiently described the process for defining and identifying the target populations of the survey, as well as developing national sampling plans. As for the creation of a restricted school frame and of randomized school IDs, 33 NRCs felt this was sufficiently described in the SOP.

According to the NRCs, there were some conditions or organizational constraints that required deviations from the standard ICILS 2023 main survey within-school sampling design in seven cases. These included the exclusion of private schools or special classes, differences between laws in different parts of a country or the overlap with a Trends in International Mathematics and Science Study (TIMSS) 2023 sample. Problems with participating in TIMSS 2023 grade 8 with the same schools were reported by six NRCs, resulting in the same students being tested in both surveys in some countries, while the overlap was intentionally avoided in others.

Table 9.17: Sampling (number of NRC responses, max 35)

	Yes	No	No Answer/ Not Applicable
Did you feel well supported by the IEA sampling team for all sampling related matters?	32	0	3
Did the Manual "Survey Operations Procedures Unit 1, Sampling Schools" sufficiently describe the following procedures?			
<i>Defining and identifying the target populations of the survey</i>	32	1	2
<i>Developing national sampling plans (exclusions, stratification, sample sizes)</i>	32	1	2
<i>Creation of a restricted school frame and of randomized school IDs?</i>	33	0	2
Were there any conditions or organizational constraints that required deviations from the standard ICILS 2023 main survey within-school sampling design?	7	26	2
If you participate also in TIMSS 2023 grade 8 with the same school: did you face any problems?	6	25	4
Did you let staff of the sampled schools complete the listing and tracking forms for teachers and students provided by the Within-School Sampling Software (WinW3S)?	25	8	2
Did you use other means instead of names to identify students and/or teachers on the forms and labels due to data protection/confidentiality laws or rules in your country?	14	19	2

Twenty-five NRCs reported they let the staff of the sampled schools complete the listing and tracking forms for teachers and students provided by the Within-School Sampling Software (WinW3S). Other means mentioned of compiling the information were using Excel sheets or internal software, and this was sometimes done by external test administrators or the NSC rather than the school staff. Concerning data protection/confidentiality laws or rules, 14 NRCs indicated they used other means instead of names to identify students and/or teachers on the forms and labels for their sake, mostly pseudonyms/code names or IDs.

Regarding the perceived difficulty of sampling related procedures, the NRC responses are shown in Table 9.18.

Table 9.18: Sampling - Procedure difficulty (number of NRC responses, max 35)

	Not difficult at all	Somewhat difficult	Very difficult
How difficult was it to...			
...adapt the international sampling design to your national specifications?	29	4	0
...create a sampling frame (i.e., the list of all eligible schools)?	30	2	1
...work with the restricted school sampling frame?	33	0	2
...select the class and teacher samples with the WinW3S software?	29	3	1

Contacting schools and recruiting school coordinators

The following section of the SAQ collected information regarding school recruitment. Table 9.19 provides the NRCs' responses to questions about school participation. Generally, countries used a variety of means to contact schools and different materials to request school participation in the main survey, including sample letters provided in the Survey Operations Procedures Manual and letters from relevant ministries. Schools' participation in ICILS 2023 was compulsory in 15 of the participating countries and benchmarking participants (for all schools in 10 countries and in some schools for the remaining five). Relatedly, the NRCs reported difficulties in convincing schools to participate in 21 of the countries. This was mostly due to the busy workload during the time of the year and limited staff capacity.

Table 9.19: Contacting schools and recruiting school coordinators (number of NRC responses, max 35)

	Yes	No/No Answer	Not Applicable
How did you contact the sampled schools?			
By phone	25	8	2
By regular mail	8	25	2
By email	32	1	2
Other	9	24	2
Which materials did you use to request school participation?			
Example letters provided in Appendix A of the Survey Operations Procedures Unit 2	11	24	0
Letters based on other national projects	12	23	0
Other	10	15	0
Is schools' participation in ICILS 2023 in your country compulsory?	15	18	2
Did you have any difficulties in convincing schools to participate?	21	12	2

School coordinators were mostly selected by the school principals/heads and trained through a variety of methods (see Table 9.20). Some school coordinators reported difficulties with understanding certain aspects of ICILS administrations to NRCs, particularly flagging students to be excluded prior to student form assignment and the necessity for listing all target grade teachers and all classes for the target grade.

Table 9.20: Training school coordinators (number of NRC responses, max 35)

	Yes	No/ No Answer	Not Applicable
How did you train the school coordinators?			
Formal training sessions	20	13	2
Through telephone, email, or video-link	19	12	2
Written instructions	25	8	2
Other	5	26	2
Did the school coordinators report difficulties with understanding any of the following aspects of ICILS administration?			
Identifying eligible teacher and/or students	6	27	2
The necessity for listing all target grade teachers	8	25	2
The necessity for listing all classes of the target grade	8	25	2
Flagging classes to be excluded prior to class sampling	3	30	2
Flagging students to be excluded prior to student form assignment	10	23	2
The processes for running the USB/internet-based delivery compatibility tests in schools	7	26	2
The steps to set up computers in schools to support successful test administration	4	29	2
The test administration procedures	5	28	2
Other	1	32	2

Adapting and translating the ICILS assessment materials

NRCs also provided information about the process of adapting and translating the ICILS assessment materials into national languages. Most NRCs identified national center staff as the individuals responsible for adapting and/or translating the survey instruments, as well as their review (see Table 9.21). In little less than half (15) of all participating countries, an outside specialist was involved in the translation (or its review in seven countries) in addition to the own staff.

Table 9.21: Adapting and translating the ICILS assessment materials (number of NRC responses, max 35)

	Adaptation	Translation
Who adapted and translated the international version of the ICILS 2023 instruments?		
Own staff	31	27
Outside translator(s)	4	15
Who reviewed adaptations and translations of the instruments (student test and questionnaires)?		
Own staff	32	30
Outside translator(s)	3	7

Most NRCs reported they were able to identify and employ a translator fluent in English (29), familiar with survey translations (29) and with a good level of computer literacy (27). However, translators in eight of the reported cases did not have the survey language as their mother tongue (Table 9.22). Similarly, while the majority of NRCs were able to find and use a reviewer fluent in English (30), familiar with the subject area and with the education in the target grade (28) and with a good level of computer literacy (28), in only 23 cases did they also have the survey language as mother tongue.

Table 9.22: Translator/reviewer competencies (number of NRC responses, max 35)

	Yes	No
Were you able to identify and employ a translator who:		
<i>Is fluent in English</i>	29	2
<i>Has the survey language as mother tongue</i>	23	8
<i>Has a good level of computer literacy</i>	27	4
<i>Is familiar with survey translations</i>	29	2
Were you able to identify and employ a reviewer who:		
<i>Is fluent in English</i>	30	1
<i>Has the survey language as mother tongue</i>	23	7
<i>Has a good level of computer literacy</i>	28	3
<i>Is familiar with the subject area</i>	28	3
<i>Is familiar with the education in the target grade</i>	28	3

A majority of NRCs did not encounter difficulties with the adaptation and/or translation of the international versions of the ICT coordinator questionnaire (31), school principal questionnaire (31), teacher questionnaire (30) or student questionnaire (29). There were more difficulties reported with the adaptation and/or translation of the student modules—seven cases for CIL and five for CT (Table 9.23). These were most commonly attributed to a lack of adaptation and translation notes, unclear or inconsistent source version, or errors in the source or National Adaptation Forms (NAF). In addition, NRCs mentioned issues with the assessment platform and trend translations.

Most countries adapted/translated the international version of the instruments directly from the English source version (32), 11 using the 2018 national version for trend items, with some also using the 2013 national version for trend items (9) or the translation provided by other national center(s) (7).

Other than translation and adapting the URL, seven NRCs reported making adaptations to the international cover letters for online administration, such as including national symbols, links to a promotional video, or additional contacts for any questions.

Table 9.23: Making and documenting national adaptations using the national adaptations form (number of NRC responses, max 35)

	Yes	No
Did you experience difficulties with the adaptation and/or translation of the international versions of the following instruments to your national context?		
<i>Student test modules CIL</i>	7	25
<i>Student test modules CT, if applicable</i>	5	21
<i>Student questionnaire</i>	4	29
<i>Teacher questionnaire</i>	2	30
<i>School principal questionnaire</i>	1	31
<i>ICT coordinator questionnaire</i>	1	31
Other than translation and adapting the URL, did you have to make any adaptations to the international cover letters for online administration?	7	26
Was the information provided in the SOP 3, Chapter 2 (Making structural adaptations) clear and sufficient for listing the structural adaptations in the NAF?	30	3

Most (30) NRCs felt the information provided in the SOP 3, Chapter 2 (Making structural adaptations) was clear and sufficient for listing structural adaptations in the NAF. The remaining three expressed a wish for clearer listing of instructions regarding the numbering adjustments after making (structural)

adaptations or adding national questions.

As for the difficulty of the procedures (Table 9.24), most NRCs expressed it was not difficult at all to include structural national adaptations in the NAF (27) or translate the ICILS adult questionnaires into their national language(s) and context (22). When it came to using the AM system to translate the instruments, a little more than half (17) of the NRCs found this not difficult at all, while others found it somewhat difficult (15) or very difficult (1). Similarly, half of the respondents (16) found it somewhat difficult to translate the ICILS student instruments into their national language(s) and context, with the other half did not find this difficult at all.

Table 9.24: Adapting and translating the ICILS assessment materials - Procedure difficulty (number of NRC responses, max 35)

	Not difficult at all	Somewhat difficult	Very difficult
How difficult was it to...			
...include your structural national adaptations in the NAF?	27	6	0
...use the AM system to translate the instruments?	17	15	1
...translate the ICILS student instruments into your national language(s) and context?	16	16	0
...translate the ICILS adult questionnaires into your national language(s) and context?	22	10	0

When asked about their experiences with the external adaptation and translation verification processes organized by the International Study Center (ISC) at IEA Hamburg and IEA Amsterdam (respectively), only a very small number of NRCs indicated that they had problems (Table 9.25), mainly with delays associated with the use of the RM AM platform (for student instruments) and communication related to that (see further information about issues with translation and adaptation processes in Chapter 5).

As for the materials that did not require verification (the *School Coordinator Manual*, the *Test Administrator Manual*, and/or the *Scoring Guides*), nearly all NRCs encountered no problems translating those.

Table 9.25: Adapting and translating the ICILS assessment materials - Issues encountered (number of NRC responses, max 35)

	Yes	No
Did you have major problems regarding the processes of verification for any of the instruments (student instruments or adult questionnaires)?		
Adaptation verification coordinated by the ISC	2	31
Translation verification coordinated by ISC	1	32
Did you have any problems with translating the <i>School Coordinator Manual</i> , the <i>Test Administrator Manual</i> , and/or the <i>Scoring Guides</i> ?		
School Coordinator Manual	0	33
Test Administrator Manual	1	32
Scoring guides	1	32

Finalizing the ICILS assessment materials

As final steps, USB sticks were reportedly copied in house (11), externally (5), or using both methods (2) where applicable. There were no security breaches reported for any of the instruments (Table 9.26).



Table 9.26: Preparing the ICILS materials for administration - Issues encountered (number of NRC responses, max 35)

	Yes	No
Did you discover any potential breaches of security for any of the instruments (student, teacher, ICT coordinator, principal)?	0	32
Did you have major problems regarding the verification process for your adult questionnaires in the IEA Online SurveySystem (OSS)?	2	30
During the administration of the adult questionnaires, were there any problems concerning the Web-Application?	12	20
During the administration of the adult questionnaires, were there any problems you observed or that were reported to you concerning the login procedures (e.g., respondents who confused ID and checksum, problems entering the URL)?	9	22

Only two NRCs reported major problems regarding the verification process for adult questionnaires in the OSS. Both of these cases (involving skipping questions) were solved in cooperation with the ISC. There were, however, problems concerning the Web-Application in 12 cases. The NRCs mentioned a system outage lasting a few days, as well as a ransomware attack (please refer to [Chapter 3](#) for more information). In addition, there were some minor issues reported concerning the login procedures (9 cases), such as difficulties accessing the correct website, opening the survey on specific devices or confusing the ID and checksum.

Administering the student instruments and adult questionnaires

The section of the SAQ inquiring about main survey administration addressed some important aspects of the data collection; selected results are presented in Tables 9.27, 9.28, and 9.29. In line with the data collected by the test administrator interviews, NRCs confirmed that at least some of their test administrators were drawn from national center staff or contracted externally, but it was more common to recruit test administrators from among the school coordinators or teachers, either of the sampled schools or from other schools. Test administrators were mostly trained through formal training sessions (29 cases), through telephone, email or video-link (16), or with written instructions (19). Other means mentioned were trainings through an external provider, helpdesk, or weekly memos.

Table 9.27: Administering the main survey - Test administrators (number of NRC responses, max 35)

	Yes	No/No Answer
Who were the test administrators for the ICILS main survey?		
National center staff	10	25
Regional or district government staff	2	33
External contractor staff	10	25
Teachers from other schools	1	34
Teachers from the sampled schools but not of the sampled class(es)	17	18
Teachers of the sampled class(es)	4	31
School coordinators	11	24
Other	4	31
How did you train the test administrators?		
Formal training sessions (face-to-face or remotely)	29	6
Through telephone, email, or video-link	16	19
Written instructions (in addition to the Test Administrator Manual)	19	16
Other	3	32

Regarding participation (Table 9.28), there were some difficulties to achieve high participation rates of teachers (in 24 countries), students (in 13 countries), ICT coordinators (in 11 countries), or principals (in 10 countries). According to the NRC responses, the participation rates were negatively impacted by several factors including the voluntary nature of the study, technical issues, and logistical challenges. Teachers were at times less motivated due to the time-consuming nature of the survey, lack of remuneration, and skepticism about the assessments. Competing international and national exams, teacher shortages, heavy workloads, and end-of-year schedules further reduced participation. Efforts like constant communication and individualized follow-ups were necessary but not always effective in overcoming these barriers.

Table 9.28: Administering the main survey - Participation issues (number of NRC responses, max 35)

	Yes	No
Did you experience any difficulties to achieve high participation rates of students, teachers, ICT coordinators or school principals within the school?		
<i>Students</i>	13	19
<i>Teachers</i>	24	8
<i>ICT coordinators</i>	11	21
<i>Principals</i>	10	21
Was the information provided in the manuals regarding the implementation of accommodations for qualified students clearly explained?		
<i>Survey Operations Procedures Manual 4</i>	28	3
<i>School Coordinator Manual</i>	28	3
<i>Test Administrator Manual</i>	29	2
Was the set of listed accommodations sufficient for the needs of students in your country?	27	5

Most NRCs thought the information provided in the manuals (SOP, school coordinator, and test administrator) regarding the implementation of accommodations for qualified students was clearly explained and that the set of listed accommodations was sufficient for the needs of students in their country. NRC responses regarding the difficulty to manage the different steps in the process of managing accommodations for qualified students are presented in Table 9.29, with keeping records of the additional time provided to students who qualified for and made use of this accommodation proving the most difficult for countries to manage.

Table 9.29: Administering the main survey - Accommodations (number of NRC responses, max 35)

	Not difficult at all	A little difficult	Somewhat difficult	Very difficult
How difficult was it to manage the different steps in the process of managing accommodations for qualified students? (if applicable)				
Having the school coordinators provide details of students qualified to receive accommodations on the student listing form.	11	5	4	1
Confirming with school coordinators that nominated students were qualified to make use of accommodations.	1	6	3	1
Using the accommodation codes on the student listing and tracking forms.	11	7	4	0
Making arrangements with schools and test administrators regarding the implementation of listed accommodations on the days of testing.	10	4	6	1
Making use of the function to disable the timer (where appropriate) on the days of testing.	13	2	6	0
Keeping records of the additional time provided to students who qualified for and made use of this accommodation.	10	3	2	6

Finally, most (28) NRCs noted that more than 75 percent of student tests and questionnaires were returned on time. As for teacher questionnaires, NRCs reported more than 75 percent of these being returned on time in 14 countries, 51–75 percent in ten countries, 26–50 percent in five countries and 25 percent or less in three countries. For ICT coordinator questionnaires, the proportion reportedly returned on time was more than 75 percent in 19 countries, 51–75 percent in seven countries and 26–50 percent in six countries. Lastly, for the principal questionnaire the proportion NRCs reported returned on time was more than 75 percent in 20 countries, 51–75 percent in five countries and 26–50 percent in seven countries.

National quality monitoring program

IEA provided materials to assist national centers in conducting their own national quality control program. The ICILS 2023 Survey Operations Procedures Unit 4 provided guidelines for selecting the national quality control monitors (NQCMS), and basic information about their duties. In addition, the ICILS 2023 *National Quality Control Monitor Manual* was developed for use in training NQCMS. The NQCMS' responsibilities were similar to those of the IQCMS, and involved visiting selected schools to observe and document the data collection sessions and interview school coordinators. Twenty-eight NRCs confirmed that they had implemented a national quality control program (Table 9.30), the number of NQCMS appointed ranging from one to 30 and the number of school visits across the participating countries being 664. In the countries where this was not implemented, it was often mentioned that test administrators were contracted from the NSC (thus deeming additional control monitoring unnecessary) or that there was a lack of time and resources.

Table 9.30: National quality monitoring program (number of NRC responses, max 35)

	Yes	No
Did you organize a National Quality Monitoring Program?	28	6
How did you train the NQCMs?		
Formal training sessions (face-to-face or remotely)	22	13
Through telephone, email or video-link	10	25
Written instructions (in addition to the NQCM Manual)	12	23
Other	4	31
Did you use the <i>National Quality Control Manual</i> and the Classroom Observation Record provided by the ISC to conduct your National Quality Monitoring Program?		
<i>National Quality Control Observer Manual</i>	28	6
<i>Classroom Observation Record</i>	27	8
Did you make any adaptations to the NQCM Manual?	6	29
Did your NQCMs document any major problems or special circumstances that occurred frequently during the assessment administration?	9	26
Did you undertake any actions as a follow-up to the reported problems (if applicable)?	6	5

NQCMs were trained mostly through formal training sessions (in 22 countries), written instructions (in 12 countries) or through telephone, email, or video link (in 10 countries). The NQCM Manual was used in all 28 countries which conducted the National Quality Monitoring Program and the COR in nearly all (27) of them. Adaptations to the NQCM Manual were made in six countries.

In line with the IQCM reports, the NQCMs confirmed the quality of the survey process was generally good. In nine countries NQCMs documented some major problems or special circumstances that occurred frequently during the assessment administration. The problems reported included test administrators not duly following the script or other administration procedures, students skipping through the assessment without making any effort, unsuitable rooms being used for the assessment or technical difficulties. In six countries, actions were undertaken by the NRC as a follow-up to these problems. In addition to NQCMs, the NRCs also collaborated with IQCMs. Only one NRC reported experiencing issues in working with the IQCM, explaining that preparing materials needed for the IQCM was cumbersome and time-consuming. Nonetheless, no other complaints were made by the NRCs regarding the appointed IQCMs.

Scoring open-ended response items and coding occupation data

NRCs were also asked to comment on the persons responsible for scoring open-ended response items, the procedures used to verify scoring reliability, and the coding of occupation data (see Table 9.31). The participating countries used scorers with a variety of different backgrounds, including national center staff, teachers/professional educators, university students or externals with degrees in relevant areas. In one country, this was done by the test administrators. The number of scorers employed in countries ranged from four to 53 and the number of days for their training, between one and 10 days.

When asked about difficulties using the web-based administration and scoring system to create user accounts for scorers and to allocate them to teams, eight NRCs reported experiencing issues with this. These difficulties were mostly related to user role permissions in the system and problems with assigning scorers to multiple teams. Five NRCs also reported difficulties using the web-based scoring system to select, score and allocate example pieces of work to be used in the scorer training. As for the scorers themselves, they experienced difficulties using the scoring system for training and then for scoring student work in 12 countries, as reported by NRCs. These included software malfunctions, such as scores not being saved or the system crashing, or the inability to assign a predetermined number



of responses to score to each user or double score additional items. However, the most frequent complaint by far was the system being slow. Only one NRC reported difficulties with the reports for reliability scoring, related to the reports being updated too late.

Table 9.31: Scoring CIL and CT test responses (number of NRC responses, max 35)

	Yes	No/No Answer
Who primarily scored the student test constructed response and performance task items?		
<i>National center staff</i>	12	23
<i>Teachers/professional educators</i>	20	15
<i>University students</i>	12	23
<i>ISC</i>	0	35
<i>Other</i>	2	33
Did you have any difficulties using the web-based administration and scoring system to create user accounts for scorers and to allocate them to teams?	8	27
Did you have any difficulties using the web-based scoring system to select, score, and allocate example pieces of work to be used in the scorer training? (if applicable)	5	19
Did the scorers (including the scoring administrator and lead scorers) have any difficulties using the scoring system for training and then for scoring student work?	12	23
Did you have any difficulties with the reports for reliability scoring?	1	34

As shown in Table 9.32, the countries mostly relied on staff from the national centers to code occupational data (23 cases), in addition to external organizations (9 cases), or university students (7 cases). The coders had experience in this area in about 60 percent (21) of the countries. In addition, NRCs had access to expert advice on occupational coding in their country in 14 countries. Countries used anywhere between two and 20 coders to code the occupation data and used a job classification scheme other than the ISCO-08 for coding occupations in three cases, in which national classification schemes were used. There were no difficulties reported with converting national coding scheme to ISCO-08 (which was available in the national language in most (26) countries), but there were some difficulties with coding student responses (7 cases) and the occupation coding sheets in Excel (6 cases).



Table 9.32: Entering and coding occupation data (number of NRC responses, max 35)

	Yes	No/No Answer
Did you use your own staff or outside experts to code occupations?		
Own staff	23	12
External organization	9	26
University students	7	28
ISC	1	34
Other	5	30
Did the coders of occupations at your national center have any experience in this area?	21	14
Did you have access to expert advice on occupational coding in your country?	14	21
Did you use any other job classification scheme than the ISCO-08 for coding occupations?	3	32
Did you have any difficulties with the following aspects of coding the occupation data?		
Converting national coding scheme to ISCO-08	0	35
Coding student responses	7	28
The occupation coding sheets in Excel	6	29
Other	3	32

Collecting and submitting data

The final data from countries was submitted to the ISC using the ICILS 2023 SharePoint site. Only one country (see Table 9.33) reported experiencing difficulties using the site when submitting and/or downloading any files (e.g., files for translation verification, USB player exchange or final data submission). It was mentioned that sometimes there were no notifications for updates and uploads and that some folders were named too generically.

Table 9.33: Collecting and submitting data (number of NRC responses, max 35)

	Yes	No
Did you experience any difficulties using the ICILS 2023 SharePoint when submitting and/or downloading any files (e.g., files for translation verification, USB player exchange, or final data submission to the ISC)?	1	34

The NRCs' assessment of the quality of the ICILS manuals and study resources

The SAQ also provided an opportunity for NRCs to give feedback on the quality of the various ICILS manuals and study resources. The NRCs were generally very positive about the manuals and webinars, and nearly all of them described these resources as "very" or "somewhat" helpful for carrying out the survey (Table 9.34). When asked for their suggestions for improvement, some NRCs suggested dividing some of the manuals in parts, highlighting important information better, or cutting down on repetitive content. More details on exclusion and accommodation or a WinW3S manual would have also been appreciated by some NRCs.



Table 9.34: The NRCs' assessment of the quality of the ICILS manuals (number of NRC responses, max 35)

	Very helpful	Somewhat helpful	A little helpful	Not helpful at all
Did you find the manuals helpful for carrying out the ICILS 2023 main survey?				
Survey Operations Procedures Unit 1 (Sampling Schools)	31	1	0	0
Survey Operations Procedures Unit 2 (Working with Schools)	27	3	2	0
Survey Operations Procedures Unit 3 (Preparing National Study Instruments)	29	3	0	0
Assessment Master Manual	28	4	0	0
Survey Operation Procedures 3 – OSS Supplement	27	5	0	0
Survey Operations Procedures Unit 4 (Data Collection and Quality Monitoring Procedures)	29	2	1	0
Survey Operations Procedures Unit 5.1	28	3	1	0
Survey Operations Procedures Unit 5.2	28	3	1	0
Preparing Computers for ICILS	26	6	0	0
School Coordinator Manual	27	5	0	0
Test Administrator Manual	29	3	0	0
Scorer training Manuals	24	8	0	0
National Quality Control Monitor Manual	23	7	0	0

Chapter 10:

Data management and creation of the ICILS database

Tim Daniel

10.1 Introduction

This chapter describes the procedures for checking the ICILS 2023 data and database creation that were implemented by IEA and the national study centers of the participating countries. For readability, this chapter refers to the principal, ICT coordinator, and teacher questionnaires as “adult questionnaires.”

Preparing the ICILS 2023 international database (IDB) and ensuring its integrity was a complex endeavor requiring extensive collaboration between the international study center (ISC) at IEA and the national study centers. The IDB was one of the key products released at the end of the study cycle as it contained all data from all participating countries. These data were the ones used by the ISC to produce the international report and all accompanying release material. Once the countries had collected the data files and submitted them to the ISC, data cleaning began. Data cleaning is an extensive process of checking data for inconsistencies and formatting the data to create a standardized output. The main goals of the data cleaning process were to ensure:

- All information in the database conformed to the internationally defined data structure;
- The content of all metadata and documentation appropriately reflected national adaptations to questionnaires;
- All variables used for international comparisons were comparable across countries (after harmonization where necessary); and
- All stakeholders involved in this process applied quality control measures throughout, to assure the quality and accuracy of the ICILS 2023 data.

10.2 Data sources

Computer-based student assessment

As a computer-based assessment, ICILS 2023 exclusively generated electronic data from the student assessment along with paper or electronic tracking information. One version of the student assessment software existed per country and administered language. Overall, 50 language versions across 32 countries were created. It included all test modules and the context questionnaire, and supported all assessment languages. The student assessment was conducted using the RM Assessment Master Examiner (see [Chapter 3](#)).

Each student was assigned an instrument version that contained two of the seven computer and information literacy (CIL) test modules as well as the student questionnaire. In those countries where the computational thinking (CT) modules were administered, students were assigned two additional CT assessment modules after the questionnaire session. All three components were administered as timed tests (see [Chapter 1](#)) and were to be done in one go by students. Short breaks in between each component were given. Administration modes offered were:

- Internet (browser-based) delivery;
- USB flash drive delivery (for use with school computers running Windows operating systems);



and

- Server mode delivery (where one school laptop hosted the software and student computers connected to this local server then).

As the default administration methods for ICILS 2023, the national study centers provided schools with one USB stick per student (plus spare sticks in case of technical failure) or the web address on which the online version of the assessment was hosted. If the computers at a school were deemed unsuitable for administering the student assessment on USB or internet connectivity was not given, a laptop computer could be used to host the application and connect school computers with it (server mode). Some countries also carried-in laptops from the national study center to overcome issues with school computers. The national study centers also provided schools with a student tracking form that notified the test administrators of the students who were to be assigned the assessment and questionnaire. The form also allowed administrators to indicate student participation for subsequent entry and verification of it in WinW3S.

In schools where one of the default administration methods was utilized, data from the student assessment were either stored on the individual USB sticks used to deliver the assessment or directly saved on an international server if the assessment was taken online. If the server method was utilized, then the student data were stored on the server laptop that was used to administer the assessment. After the assessment, the national study centers uploaded the data from USB sticks or server laptop to a central international server. For assessments conducted using the online method, an upload was obsolete.

The data were transposed from long format to a wide format so that they could be transformed for processing and analysis. This process resulted in a predefined table structure, containing one record per student as well as all variables that were to be used for data processing, analysis, and reporting. During this step, a set of calculations needed to take place. Examples include automatic scoring of some of the complex or constructed-response items and large tasks or aggregating time spent on an item across multiple visits to it.

Online data collection of adult questionnaires

ICILS 2023 was offered as an online-only data collection for the adult questionnaire. All participating countries used the OSS software to administer the adult questionnaires.

To ensure confidentiality, national study centers provided every respondent with a letter that contained individual login information along with information on how to access the online questionnaire. This login information corresponded to the respondent identification (ID) and password provided from the WinW3S software, meaning that the identity validation step occurred when the respondents' logged-in to the survey.

As respondents completed their online questionnaires, their data were automatically stored on a central international server. Respondents could log in multiple times until they actively submitted their responses and, with this, indicated that they finalized it. Once that occurred, the questionnaire was locked for this respondent from then on. Data for each country-language combination were stored in separate tables on the server. The different language versions within countries were then merged (at the ISC) with the data collected as part of the within-school sampling process in WinW3S (for details, please refer to [Chapter 8](#)).

Data verification at the national study centers

Before sending the data to the ISC for further processing, national study centers were to carry out mandatory validation and verification steps on all data and apply corrections as necessary. The corresponding routines were included in the WinW3S software, which automatically and systematically checked ID codes present in the OSS and AM Examiner databases to correspond to the participation indicator variables and entries on the tracking forms and as entered in WinW3S. The ID codes

were imported into WinW3S from Excel files formerly created from the OSS and AM systems for this purpose.

In addition to submitting the data files described above, national study centers provided the ISC with detailed data documentation, including hard copies or electronic scans of all original, filled student and teacher tracking forms. The ISC already had access, as part of the layout verification process, to electronic copies of the national versions of all questionnaires and the final national adaptation forms.

While the questionnaire data were collected, each national study center used the information from the student and teacher tracking forms to verify the completeness of the materials. Participation information (e.g., whether the concerned teacher or student had left the school permanently between the time of sampling and the time of administration) was entered via the WinW3S software.

This process was also supported by the option in WinW3S to generate an inconsistency report. This report listed all discrepancies between variables recorded during the within-school sampling and test administration process and so made it possible to cross check these data against the actual availability of data entered in the online questionnaires, and the uploaded student data on the central international server hosted by RM. Data managers were requested to resolve these problems before final data submission to the ISC. If inconsistencies had to remain or the national study center could not solve them, the ISC asked the center to provide documentation on these problems. The ISC used this documentation when processing the data at a later stage.

10.3 Confirming the integrity of the national databases

Overview

As described earlier in this chapter, national study centers in each participating country were responsible for verifying their national ICILS 2023 data and submitting all accompanying documentation as well as a finalized WinW3S database including final participation information to the ISC. Furthermore, the data from the adult online questionnaires were automatically stored on a central international server hosted by IEA, and the results from the student assessment on a central server hosted by RM. Hence, the ISC already had the adult questionnaire data available on its server and downloaded the student data from RM hosted servers once a country announced completion of data collection. Data from both data sources were then merged together with the participation information received via the WinW3S database and fed into a comprehensive processing system for checking and editing. To facilitate the data cleaning process, the ISC asked the national study centers to provide them with detailed documentation of their data together with their national data files. The data documentation included copies of all original survey tracking forms, as well as any additional information from field operations within a country provided by the national study center. They also submitted their final national adaptation forms to provide and confirm complete documentation on all national adaptations. In addition, national study centers were asked to provide documentation on all changes or edits applied to the data prior to submission, as well as any issues remaining in the data which were already verified and deemed that they could remain.

Ensuring the integrity of the international database required close cooperation between the international and national institutions involved in ICILS 2023. After each country had submitted its data and required documentation, the ISC, in collaboration with the national research coordinators (NRCs), conducted several cleaning procedures upon these:

1. Checking documentation, import, and structure;
2. Cleaning identification (ID) variables;
3. Checking linkages; and
4. Resolving inconsistencies in questionnaire data.

The cleaning process was an iterative process. Numerous iterations of the cleaning procedure were

completed on each national data set. This repetition ensured that all data were properly cleaned and that any new errors that could have been introduced during the data cleaning were rectified. The cleaning process was repeated as many times as necessary until all data were consistent and comparable. Any inconsistencies detected during the cleaning process were resolved in collaboration with national study centers, and all corrections made during the cleaning process were documented in a cleaning report produced for each country.

Having completed this work, the ISC staff sent queries to the national study centers where inconsistencies were found. These required the centers to either confirm the ISC's proposed data-editing actions or provide additional information to resolve inconsistencies. After all modifications had been applied, the ISC rechecked all datasets. This process of editing the data, checking the reports, and implementing corrections was repeated as many times as necessary to help ensure that data were consistent within and comparable across countries.

After the national data had been checked, the ISC provided national study centers with univariate statistics. This material enabled national study centers to compare their national data with the international results as they were included in the draft international report and related data and documentation.

This step was one of the most important quality measures implemented, because it helped to ensure to increase validity of the data. For example, a particular statistic that might have seemed plausible within a national context could have appeared as an outlier when the national results were compared with the international results. The outlier could hint to an error in translation, data capture, coding, etc. The team at the ISC reviewed all such instances and, where necessary, addressed them, for example, by recoding the corresponding variables in appropriate ways or, if errors could not be corrected, removing them from the international database.

Once the national databases had been verified and formatted according to the international file format, the ISC produced data files containing information on the participation status of students, teachers, principals, and ICT coordinators in each country's sample. The ISC then used this information to calculate sampling weights, population coverage, and school, teacher, and student participation rates. [Chapter 7](#) of this report provides details about the weighting procedures.

In a subsequent step, the IEA analysis team estimated CIL performance and CT scores as well as questionnaire indices for students, teachers, and school-level survey data (see [Chapter 11](#) and [Chapter 12](#) for scaling methods and procedures). On completing their verification of the sampling weights and scale scores, the the ISC included these derived variables in the international database and for distribution to the national study centers.

Data cleaning quality control

Because ICILS 2023 was a large and highly complex study with high standards for data quality, maintaining these standards required an extensive set of interrelated data checking and data cleaning procedures. To ensure all procedures were conducted in the correct sequence, that no special requirements were overlooked, and that the cleaning process was implemented independently of the persons in charge, the data quality control included the following steps:

- Thorough testing of all data cleaning programs: before using the programs to process real datasets, the ISC used mocked simulation datasets containing all possible problems and inconsistencies to check the accurate set up and work of the cleaning programs.
- Registering all incoming data and documents in a specific database: the ISC recorded the date of arrival as well as specific issues requiring attention.
- Carrying out data cleaning according to strict rules: deviations from the cleaning sequence were not possible, and the scope for involuntary changes to the cleaning procedures was minimal.

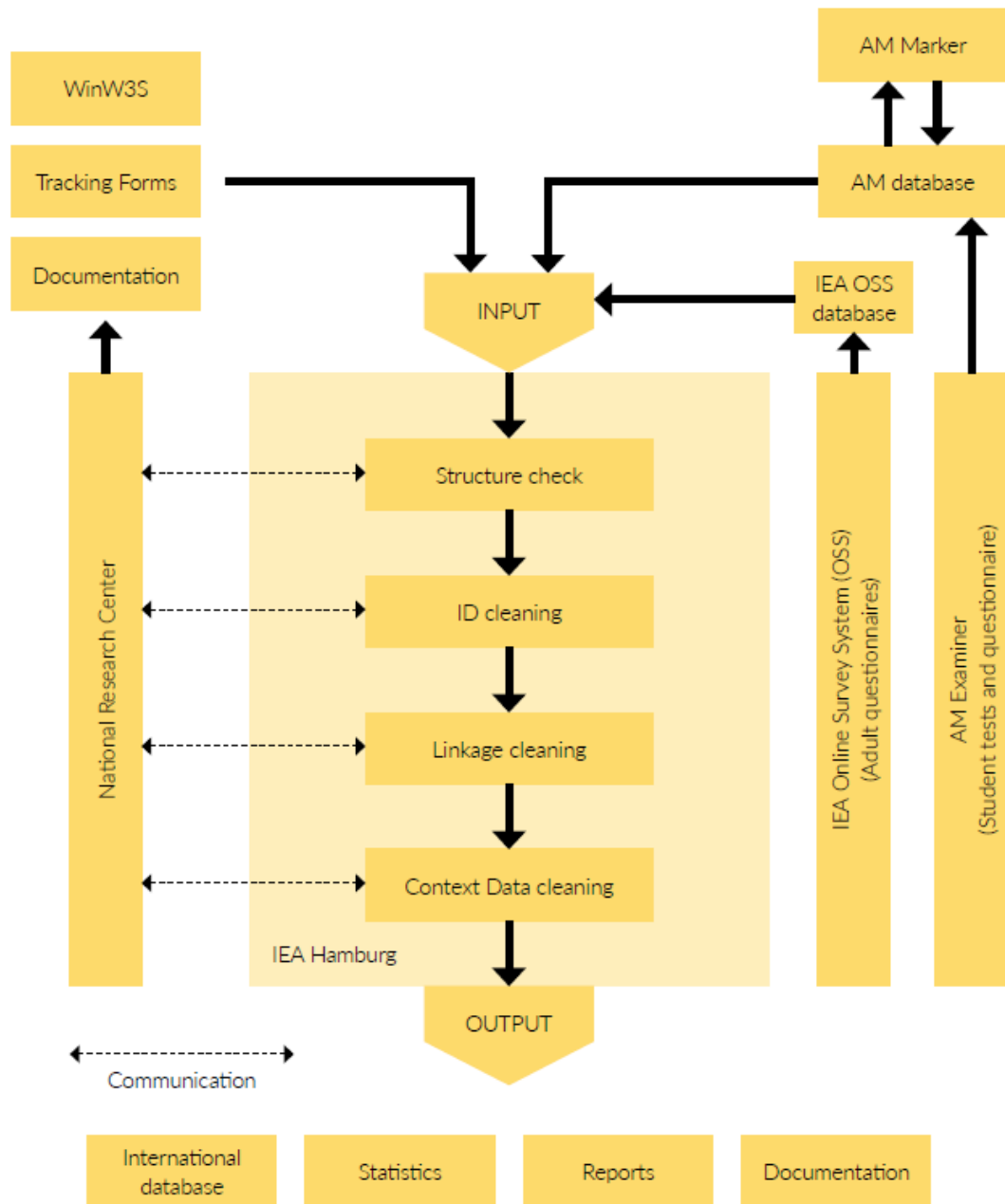
- Documenting all systematic data recodings that applied to all countries: the ISC recorded these in the ICILS 2023 Main Study General Cleaning Documentation.
- Documenting every case-level and country-specific correction to a country's data files in a recoding script: logging these changes, which only occurred occasionally, allowed the ISC to undo changes or to redo the whole manual cleaning process at any later stage of the data-cleaning process.
- Repeating, anytime and especially after completion of data cleaning for a country, all cleaning steps from the beginning: this step allowed the ISC to detect any problems that might have been inadvertently introduced during the data-cleaning process.
- Working closely with national study centers at various steps of the cleaning process: the ISC provided national study centers with the processed data files and accompanying documentation and statistics so that center staff could thoroughly review and correct any identified inconsistencies.

Preparing national data files for analysis

The main objective of the data cleaning process was to ensure that the data adhered to international formats, that student, teacher, and school-level survey data could be linked across different survey data files, and that the data reflected the information collected within each country in an accurate and consistent manner.

The program-based data cleaning consisted of the activities summarized in [Figure 10.1](#) and explained in the following subsections. The ISC carried out all these activities in close communication with the national study centers.

Figure 10.1: Overview of data processing at the ISC



Checking documentation, import, and structure

For each country, data cleaning began with an exploratory review of its data file structures and data documentation (i.e., national adaptation forms, student tracking forms, teacher tracking forms, survey activities questionnaire).

The ISC began data cleaning by combining the tracking information and sampling information captured in the WinW3S database with the student-level database containing the corresponding student survey instrument data. During this step, the ISC staff also retrieved the data from the principal, ICT coordinator, and teacher questionnaires. This step also saw data from the different sources being transformed and imported into one structured query language (SQL) database so that this information would be available during further data processing stages.

The first checks identified differences between the international and national file structures. Some countries made adaptations (such as adding national variables or omitting or modifying international variables) to their questionnaires. The extent and nature of such changes differed across countries: some countries administered the questionnaires without any modifications (apart from translations and necessary adaptations relating to culture or language-specific terms), whereas other countries inserted response categories within existing international variables or added national variables.

To keep track of adaptations, the ISC asked the national study centers to complete national adaptation forms. Where necessary, the ISC modified the structure and values of the national data files to ensure that the resulting data remained comparable across countries. Details about country-specific adaptations to the international instruments can be found in Supplement 2 of the ICILS 2023 user guide for the international database (Meyer et al., [2025](#)).

At this stage, the ISC discarded variables created purely for verification purposes during administration and made supplement provision for new variables necessary for analysis and reporting. The latter included reporting variables, derived variables, sampling weights, and scale scores.

The ISC referred problems that could not be rectified automatically to the responsible NRC so that the national study centers could check the original data collection instruments and tracking forms to trace the source of these errors. Wherever possible, the ISC suggested a remedy and asked the national study centers to either accept or propose an alternative. If a national study center could not resolve problems through verification of the instruments or forms, the ISC applied a general cleaning rule to the files to rectify the error. When all automatic updates had been applied, the ISC ran individual recodings in the data to directly apply any remaining corrections to the data files.

Cleaning identification (ID) variables

Each record in a data file needs to have a unique ID number. The existence of records with duplicate ID numbers in a file implies an error of some kind. If two records in an ICILS database shared the same ID number and contained exactly the same data, the ISC deleted one of the records and kept the other one in the database. If both records contained different data and the ISC found it impossible to identify which record contained the “true data,” they removed both records from the database. The ISC tried to keep such losses to a minimum; actual deletions were very rare.

Although the ID cleaning covered all data from all instruments, it focused mainly on the student file given ID mix-ups or duplicate use was way more likely here. This step in data cleaning included the preparation of the student test records. Due to the administration of the student test on USB sticks, data uploaded after test sessions often contained several student records within a country with the same student ID. In most of the cases, such records were duplicates. In few cases, multiple students used the same student ID and the correct ID had to be evaluated and assigned after the fact.

The possible sources of multiple records were tracked back to the test administration procedures at schools or technical constraints of the student test delivery software. Depending on the nature of a multiple session, the records were used for processing, deleted, assigned a different/corrected ID, or merged. In addition to checking the unique student ID number, it was crucial to check variables pertaining to student participation and exclusion status, as well as students’ dates of birth and dates of testing to calculate student age at the time of testing. The student tracking forms provided an important tool for resolving anomalies in the database. The records were cleaned by the ISC with confirmation from national study centers for individual records.

After national study centers had confirmed the cleaned ID variables, the clean databases with information about student participation and exclusion were used by the IEA sampling team to calculate students’ participation rates, exclusion rates, adjudication flags, and student sampling weights (see [Chapter 7](#) for details).

**Checking linkages**

In ICILS 2023, student, teacher, and school-level survey data appeared in different data files at the respective levels. Correctly linking these records to provide meaningful data for analysis and reporting was therefore vital. Linkage was implemented through a hierarchical ID numbering system as described in [Chapter 8](#) of this report. Student ID numbers in the main student data file had to be matched correctly with those in the reliability scoring file. Ensuring that teacher and student records linked to their corresponding schools was also important.

Resolving inconsistencies within questionnaire data

The amount of inconsistent and implausible responses in questionnaire data files (examples, see in section below) varied considerably among countries. The ISC determined the treatment of inconsistent responses on a question-by-question basis, using all available documentation to make an informed decision. The ISC also checked all questionnaire data for consistency across the responses given.

For example, Question 1 in the principal questionnaire asked for the total school enrollment (number of boys and number of girls, respectively) in all grades, while Question 2 asked for the enrollment in the target grade only. Clearly, the number given as a response to Question 2 could not exceed the number provided in Question 1. Similarly, it was not possible for the sum of all full-time teachers and part-time teachers, as asked in Question 4, to equal zero. In another example, the ICT coordinator questionnaire Question 7a asked for the total number of ICT devices in the school, and the number of ICT devices available to students. The total number of ICT devices in the school could not be smaller than the number available to students.

The ISC flagged inconsistencies of this kind and then asked the national study centers to review. The ISC recoded those cases that could not be corrected or where the response provided was not usable for analysis as “omitted.”

Filter questions, which appeared in some questionnaires, directed respondents to a particular sub-question or further section of the questionnaire. The ISC applied the following cleaning rule to these filter questions and the dependent questions that followed: If the answer to the filter question is “no” or “not applicable,” any responses to the dependent questions were recoded as “logically not applicable.”

The ISC also applied what is known as a split variable check to questions where the answer was coded into several variables. For example, Question 2 in the ICT coordinator questionnaire asked: “Which of the following teaching duties do you have?” ICT coordinators responses were captured in a set of four variables, each one coded as “Yes” if the corresponding “Yes” option was checked and “No” if the “No” option was filled in. Occasionally, ICT coordinators checked the “Yes” boxes but left the “No” boxes unchecked. Because, in these cases, it was clear that the unchecked boxes actually meant “No,” these responses were recoded accordingly, provided that the ICT coordinators had given affirmative responses in the other categories.

Resolving inconsistencies between questionnaire and tracking data

Two different sets of ICILS 2023 data indicated age and gender for both teachers and students. The first set was tracking information provided by the school coordinator or test administrator throughout the within-school sampling and test/questionnaire administration process. The second set comprised of actual responses given by individuals in the contextual questionnaires. In some cases, data across these two sets did not match and resolution was needed.

If the information on gender or birth year and month was missing in the student questionnaire but the student participated, then this data was copied over from the tracking information to the questionnaire, if available.

The teacher questionnaire did not ask teachers to provide birth year and month but rather to list



their age. Year of birth, which was indicated in the tracking forms, was then cross checked against the age indicated by the questionnaire responses. If gender and/or age information was missing from the teacher questionnaire but the teacher participated, this data was copied over from the tracking information to the questionnaire.

If discrepancies were found between tracking and questionnaire gender and age data, the questionnaire information (for both teachers and students) replaced the tracking information.

Note for countries who adapted questions about gender Several countries opted to adapt questions on gender and added a third category for selection. For cases that selected the third category, the information from tracking forms were **not** considered as described above. Here, the data for the international variables were eventually set to missing for these cases rather. For information on which countries adapted the gender questions, please refer to Supplement 2 of the ICILS 2023 user guide for the international database (Meyer et al., 2025).

Handling of missing data

Two types of data existed: valid data values and missing data values. Missing data can be assigned a value of “omitted or invalid,” or “not administered.” The ISC applied additional missing codes to the data to facilitate further analyses. This process led to four distinct types of missing data in the international database:

- *Presented but not answered, Omitted:* The respondent had a chance to answer the question but did not do so, leaving the corresponding item or question blank. Alternatively, the response was non-interpretable or out of range.
- *Not administered:* This signified that the item or question was not administered to the respondent, which meant that the respondent could not read and answer the question. The not administered missing code was used for those student test variables that were not in the sets of modules administered to a student either deliberately (due to the rotation of modules) or, in very few cases, due to technical failure or incorrect translations. The missing code was also used for those records that were included in the international database but did not contain a single response to one of the assigned questionnaires. This situation applied to students who participated in the student test but did not answer the student questionnaire. It also applied to school-level data where only one of the principal or ICT coordinator questionnaires was returned with responses. In addition, the not administered code was also used for individual questionnaire items that were not administered in a national context because the country removed the corresponding question from the questionnaire or because the translation was incorrect.
- *Logically not applicable:* The respondent answered a preceding filter question in a way that made the following dependent questions not applicable.
- *Not reached* (applied only to individual items of the student test): This code indicated those items where it was believed that the students did not reach because of a lack of time. This items were to be used different in scaling procedures than omitted codes then. “Not reached” codes were derived as follows: First, the last answer given by a student in a session was identified. Following, the first omitted response after this last answer is coded as ‘omitted,’ but all following responses to items in the session are then coded as ‘not reached.’ For example, the response pattern ‘1 9 4 2 9 9 9 9 9’ (where ‘9’ represents ‘omitted’) is recoded to ‘1 9 4 2 9 7 7 7 7’ (where ‘7’ represents ‘not reached’). When recoding ‘missing’ items to ‘not reached,’ all items, which were coded to ‘not administered’ have been ignored. For example, the pattern ‘3 1 5 2 9 9 9 8 9 9’ is recoded to ‘3 1 5 2 9 7 7 8 7 7’.

Checking the interim data products

Building the international database was an iterative process. Once the ISC completed each major data

processing step, it sent a new version of the data files to the national study centers so that they could review their data and run their own separate checks to validate the new data file versions. This process implied that national study centers received several versions of their data, and their data only, before release of the draft and final versions of the international databases.

The ISC sent the first version of data and accompanying documentation to national study centers on 3 November, 2023. At this time, data for all countries that administered ICILS in the first half of 2023 were sent out. The data for countries who administered ICILS in the second half of 2023 received their data on a rolling basis, i.e., when the data was fully processed and cleaned, starting 20 December, 2023. This first version of each country dataset included the following data and documentation:

- Student-, teacher, and school-level SPSS and SAS data files;
- Univariate descriptive statistics for all variables in the MS Excel files;
- A cleaning report that included a list of structural and case-level findings;
- A recoding documentation for country-specific data edits applied by the ISC;
- Cleaning documentation describing the initial cleaning procedures undertaken at the ISC and describing the data files and statistics provided; and
- A codebook containing variable specifications for all provided variables.

During this stage of the data processing procedure, national study centers were asked to review the documentation on national adaptations to their instruments and the related edits applied to the data files. Thus, the response data were finalized prior to further analyses, namely calculating sampling weights and scaling.

The ISC provided the IEA analysis team with subsequent versions of the data and related documentation as soon as it had implemented feedback from national study centers. These additional versions of the data files were accompanied with the sampling weights and international achievement scores as soon as these became available. During this stage of the data processing process, the ISC asked national study centers to review the documentation on adaptations to the national versions of their instruments and the related edits applied to the data files.

In April 2024, all national study centers received the data from all other ICILS 2023 countries. This data version is called the draft international database since it roughly reflects the structure of the released databases already but certain changes are still possible and expected. Most prominently and compared to earlier data send-outs this version included sampling weights and scales. All persons within the national study centers needed to sign a confidentiality agreement assuring no sharing of any ICILS 2023 data products or information with respect to the results.

During the fifth NRC meeting in Taipei, Taiwan in June 2024, NRCs had the opportunity to raise any further issues concerning their data that had not yet been raised. In July 2024, the ISC provided NRCs with an updated version of the draft international database. This version was necessary due to minor edits resulting from internal quality control procedures. The ISC used this version of the data to produce the updated, final tables for the international report.

10.4 The ICILS 2023 international database

The ICILS 2023 international database incorporated all national data files from participating countries. The data processing and validation at the international level helped to ensure that:

- Information coded in each variable was internationally comparable;
- National adaptations were reflected appropriately in all variables;
- Questions that were not internationally comparable were removed from the database;

- All entries in the database could be linked to the appropriate respondent—student, teacher, principal, or ICT coordinator;
- Only those records considered as participating (following adjudication) remained in the international database files;
- All variables derived by the ISC from result data were correctly added;
- Sampling weights and student achievement scores were available for international comparisons; and
- Indirect identification of individuals was prevented by applying confidentiality measures, such as scrambling ID variables or removing some of the personal data variables that were needed only during field operations and data processing.

More information about the ICILS 2023 international database is provided in the ICILS 2023 user guide for the international database (Meyer et al., 2025).

10.5 Summary

To achieve a high-quality database, ICILS 2023 implemented a series of data management procedures that included checks to ensure the consistency of national database structures, proper documentation of all national adaptations, and ensure the comparability of international variables across national datasets. All national databases were reviewed in cooperation with national study centers and the larger international team. The review process followed a series of thorough checking procedures, which led to the creation of the final ICILS 2023 database. The final data products included the following data and materials:

- The **international database**, i.e., student, teacher, and school-level data files for all countries in SPSS, SAS, and R format;
- A **user guide** on how to work with the international database using the IEA IDB Analyzer software;
- **Item almanacs** including statistics and information on all achievement items;
- A **codebook** describing all variables contained in the international database and their attributes in Excel format; and
- Information from the **national context survey** in SPSS format and an accompanying **codebook** in Excel.

References

Meyer, S., Daniel, T., & Tieck, S. (2025). ICILS 2023 user guide for the international database. International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/user-guides/icils-2023-user-guide>

Chapter 11:

Scaling procedures for ICILS 2023 test items

Yuan-Ling Liaw, Mojca Rožman, Julian Fraillon, and Eugenio Gonzalez

11.1 Introduction

This chapter describes the procedures used to scale the ICILS 2023 test items that were designed to measure students' computer and information literacy (CIL) and computational thinking (CT). The chapter begins with an overview of the scaling methodology as used in ICILS. The scaling models used for ICILS 2023 data are described, focusing on the Rasch model and its assumptions. Further it details the use of population modeling, multiple imputation or "plausible values" methodology, which enables proficiency generation optimal for group reporting of results. The overview concludes with a short description on equating that is used to report ICILS 2023 achievement scores on the existing ICILS achievement scales.

The chapter then explains the implementation of the scaling procedures in three sections:

- *Test and item adjudication* includes information about test targeting, dimensionality and test reliability, local dependence, treatment of item-level non response, item statistic summaries, scorer agreement, and ends with reporting the item adjudication outcomes.
- *Item calibration* includes a description of the procedures for calibrating the test items.
- *Population modeling to estimate students' scores* contains details on the principal component analysis, calculating the latent regression model, drawing plausible values, and transforming them for placement on ICILS scales as well as estimating the linking error.

The chapter concludes with a description of the process used to develop the the CIL and CT described proficiency levels.

In this chapter data from the computer-based assessment in 34 countries and the one German benchmarking participant are used for the purpose of main survey scaling analyses. The ICILS 2023 test instruments comprised seven CIL and four optional CT modules.

The modules are paired into a test form in a fully balanced rotation design. Each student responds to a test form containing either two CIL modules or two CIL and two CT modules, if the country opted for CT. This approach allows ICILS to assess a broad range of computer and information literacy and computational thinking skills, while minimizing the response burden on individual students (see [Chapter 2](#) for more details on the test instruments).

Because students are only administered a subset of the test instrument content, direct comparisons of group estimates using conventional statistics, such as total scores, is not appropriate. To produce achievement estimates, ICILS employs Item Response Theory (IRT), a psychometric scaling method that combines student responses, and "plausible value" methodology, which involves multiple imputations. ICILS assigns proficiency scores to each student based on both item parameters from the IRT model and parameters from latent regression models based on questionnaire data. This approach allows for the description of student performance across various groups or subgroups and estimating relationships between achievement and context variables.



11.2 Overview of scaling methodology

Scaling model and assumptions

One parameter (Rasch) item response theory model

The scaling of ICILS achievement data relies on item response theory (IRT) methods. IRT is particularly useful for developing comparable achievement estimates across students, even if they respond to different sets of items, as was the case in ICILS 2023. Each student completed two of seven available CIL modules and, where applicable, two of four available CT modules. In each case, the modules were allocated to students using fully balanced assessment designs that enable unbiased comparisons of achievement estimates based on different forms across groups (see [Chapter 2](#) for details of the assessment design).

As in ICILS 2013 and ICILS 2018, both the Rasch model (Rasch, 1960) and, its extension to accommodate partial-credit items, the Rasch partial credit model (PCM) (Masters & Wright, 1997) were used. This latent variable model captures the relationship between a student's proficiency and their likelihood of correctly answering each item (or achieving each score on a partial-credit item).

For dichotomous items, the Rasch model estimates the probability of a student correctly answering item i and receiving a score of 1 rather than 0 as follows:

$$P_i(\theta_n) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \quad (11.1)$$

where $P_i(\theta_n)$ represents the probability that person n scores 1 on item i , θ_n is the estimated latent trait of person n (i.e., ability), and δ_i is the estimated location of item i on the latent dimension (i.e., difficulty). For each item, the responses are modeled as a function of the latent trait θ_n .

For items with more than two response categories (e.g., items scored 0, 1, or 2), this model can be extended to the Rasch partial credit model, which is expressed as:

$$P_{xi}(\theta_n) = \frac{\exp(\sum_{k=0}^x (\theta_n - \delta_i + \tau_k))}{\sum_{k=0}^{m_i} \exp(\sum_{k=0}^k (\theta_n - \delta_i + \tau_k))}, \quad x = 0, 1, \dots, m_i, \quad (11.2)$$

where $P_{xi}(\theta_n)$ denotes the probability that person n scores x on item i out of the m_i possible scores. θ_n represents the person's latent trait, δ_i is the item parameter indicating the location of item i on the latent continuum, and τ_k is the step parameter for each category, with $\sum \tau_k = 0$.

Assumptions in using the Rasch model and its characteristics

As with the application of any mathematical model, the use of IRT makes assumptions about the data to which the model is applied. The test, and item evaluation processes described later in this chapter are conducted in order to test the degree to which the assumptions are met by the ICILS data.

Unidimensionality and local independence

The unidimensionality assumption posits all the items in an instrument measure a single and common latent trait (or measurement construct). In the case of the ICILS tests, there are two constructs that are measured, CIL and CT. The Rasch IRT manifestation of unidimensionality is that a single latent variable (θ) accounts for the probability of a test taker's responses to the CIL or CT items, as the two constructs were scaled separately.

The local independence assumption states that item response probabilities are not influenced by other factors, but solely depend on the latent variable and not on responses to other items within the same module or test. These assumptions were tested in previous ICILS cycles (see (Gebhardt & Schulz, 2015; Ockwell et al., 2020)), and item parameters for the CIL and CT domains were estimated using a



unidimensional IRT model.

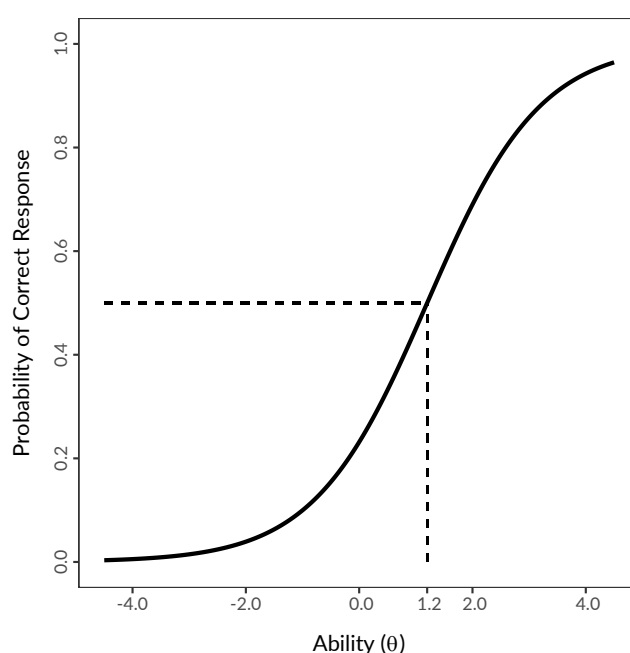
Measurement equivalence across groups

Measurement equivalence is necessary to ensure that derived scores are comparable across groups. It is achieved when the probability of an item response depends only on the person's measured latent trait, and not on their group membership. Differential item functioning (DIF) indicates that a specific item of interest is not equivalent across groups. Measurement equivalence in ICILS 2023 (as in previous cycles) was tested by gender (for the pooled international data across countries) and by country.

Monotonicity

Monotonicity, a central concept in IRT, is the assumption that when scores on the latent trait (typically referred to as *ability* in achievement tests) increase the probability of a correct response to any given item increases as well. This also reflects the intuitive idea that, when measuring an achievement construct (CIL and CT in ICILS) test-takers with higher ability are more likely to succeed when responding to the test items. When testing the assumptions of monotonicity at the item-level, we used the IRT-based infit measure of fit (Bond & Fox, 2015; Linacre, 2017), the classical test theory discrimination measure (point biserial correlation was used in ICILS), and the item characteristic curve (ICC). An ICC plots the probability of a correct response to an item as a function of the latent trait. For binary IRT model, the ICC is assumed to have an S shape. An example of the ICC is presented in Figure 11.1.

Figure 11.1: Item characteristic curve



Equal discrimination

In the Rasch IRT model, all items are assumed to discriminate equally among individuals of different abilities implying that every item is equally effective at distinguishing between test-takers with higher or lower abilities. This contrasts with other IRT models (like the 2- and 3-parameter models) that allow discrimination to vary across items. When evaluating the fit of the items to the Rasch model, we compare the model ICC for the item (noting that the shape of the model curve is the same for all items, and the horizontal position of the curve varies according to the difficulty of the item) with a plot of the probabilities based on the data for that item.



Population modeling

Balanced block instrument designs, in which individual students are systematically allocated modules and respond to a subset of items from the main item pool, is widely used in assessments like ICILS. However, reducing the test length and administering subsets of items to individual students introduces a considerable degree of uncertainty in the measurement of the latent trait at the individual level. Aggregating student ability in this way can lead to biased population estimates. ICILS uses a population model to estimate distributions of ability based on item parameters and latent regression of ability on contextual data. This provides more accurate population and subgroup estimates (Mislevy, 1991; Mislevy & Sheehan, 1987; von Davier et al., 2009).

In the latent regression model, also known as the population or conditioning model, the posterior distribution of the proficiency variable (θ) is modeled as dependent on both cognitive item responses (X) and several predictors (Y) derived from the context questionnaires. Generating plausible values requires both the item parameters from the IRT scaling stage and the estimates from the latent regression analysis.

The regression uses the context questionnaire variables to predict the proficiency variable θ (see Sub-section 11.5 for a description of how the variables are transformed and selected for the regression model). It is assumed that:

$$\theta \sim N(\mathbf{y}\Gamma, \Sigma). \quad (11.3)$$

The latent regression parameters Γ and Σ are estimated conditional on the previously determined item parameter estimates. Γ is the matrix of regression coefficients, and Σ is a common residual variance-covariance matrix.

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

$$\theta_j = \gamma_{s0} + \gamma_{s1}y_1 + \dots + \gamma_{sL}y_L + \epsilon_s, \quad (11.4)$$

where ϵ_s is an error term.

The residual variance-covariance matrix is given by the following equation:

$$\Sigma = \Theta\Theta^t - \Gamma(YY^t)\Gamma^t. \quad (11.5)$$

The conditional distribution from which plausible values for each respondent j are drawn can be written as follows:

$$P(\theta_j | x_j, y_j, \Gamma, \Sigma). \quad (11.6)$$

Using standard rules of probability, this posterior probability of proficiency can be represented as follows:

$$P(\theta_j | x_j, y_j, \Gamma, \Sigma) \propto P(x_j | \theta_j, y_j, \Gamma, \Sigma) P(\theta_j | y_j, \Gamma, \Sigma) = P(x_j | \theta_j) P(\theta_j | y_j, \Gamma, \Sigma) \quad (11.7)$$

where θ_j is a vector of scale values (these values correspond to the performance on CIL and CT), $P(x_j | \theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items



within each scale, and $P(\theta_j | y_j, \Gamma, \Sigma)$ is the multivariate joint density of proficiencies of the scales, conditional on the observed value y_j of context questionnaire responses and latent regression parameters Γ and Σ . As described above, the item parameters obtained from the IRT scaling are treated as fixed values in the estimation.

An expectation–maximization (EM) algorithm is used for estimating Γ and Σ ; the basic method for the single scale case is described in Mislevy (1985) and Mislevy (1991). The EM algorithm requires the computation of the mean and variance of the posterior distribution in the equation (11.7).

Reporting on the existing ICILS achievement scales

The metrics for the CIL and CT scales were originally established in ICILS 2013 (Gebhardt & Schulz, 2015) and ICILS 2018 (Ockwell et al., 2020), respectively. These metrics were set by normalizing the international average scores for all participating countries meeting the sampling requirements (excluding benchmarking participants) to a mean of 500 and a standard deviation of 100. So that scores can be compared across cycles, a large subset of items is reused in the next cycle and achievement data from subsequent ICILS assessment is transformed and aligned with these original metrics. This alignment is achieved using a concurrent calibration approach, which combines data from the previous cycle alongside current assessment data. Common items and common countries contribute to linking or equating the achievement scores to the established metric.

11.3 Implementation of scaling procedures: test and item adjudication

To evaluate the psychometric quality of ICILS 2023 achievement tests and items, we calculated a range of statistics for the tests and for each item in each country. These statistics were carefully reviewed as part of the ICILS quality assurance process to identify any problems with the measurement properties of the tests and items within tests. All test and item analysis as well as item calibration and plausible value imputation relied on the open-source package TAM (Robitzsch et al., 2024), available in the R statistical programming language (R Core Team, 2023).

Test targeting

As a prelude to the main IRT scaling, the preliminary Rasch analysis was conducted and the distributions of CIL and CT were estimated, respectively, along with the location of the corresponding item thresholds (with a response probability, $rp = 0.5^{20}$; see Figure 11.2 and Figure 11.3). In each figure, the left panel displays a histogram of student ability estimates, arranged from the most able students at the top to the least able at the bottom. The middle section shows the distribution of items on the same latent scale, with the most difficult items positioned at the top and the least difficult at the bottom. The right panel presents country-level ability estimates, where each country is denoted by the letter C. For dichotomous items, the item thresholds were equal to the item difficulties. In the case of partial credit items, a difficulty threshold was estimated for each score level.²¹

In both the CIL and CT assessments, the range of item difficulties generally aligned with the abilities of students in the population. However, the match between test item difficulty and student ability varied significantly across countries, reflecting differences in student achievement distributions within each ICILS 2023 country.

²⁰ This means that a respondent with the same score on the latent continuum as the item location parameter has a 50 percent probability of giving a correct response.

²¹ This “Thurstonian” threshold for a score category indicates the point on the continuum where a respondent has a 50 percent probability of obtaining this item score or higher.

Figure 11.2: Mapping of CIL student abilities, item difficulties, and country means

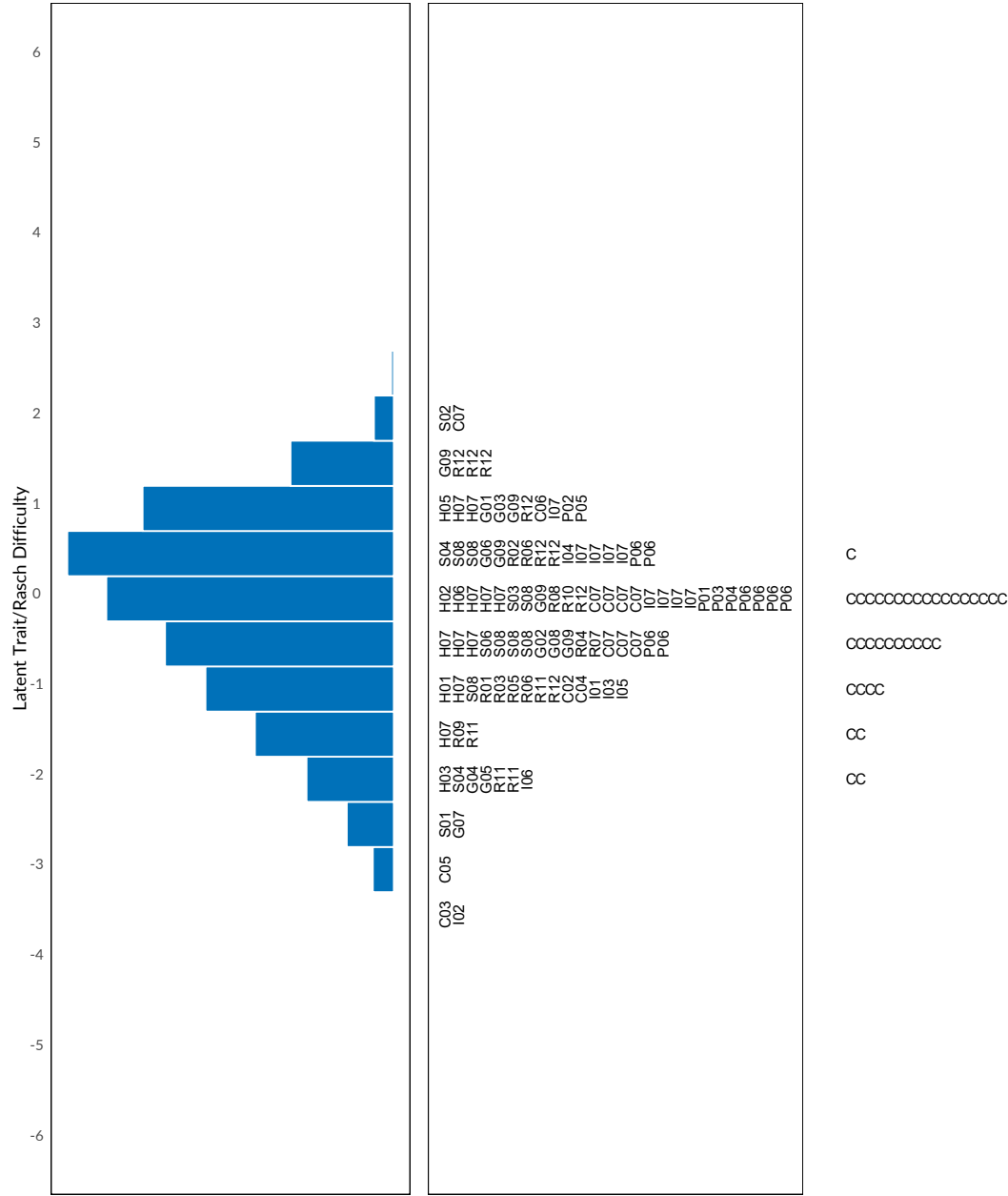
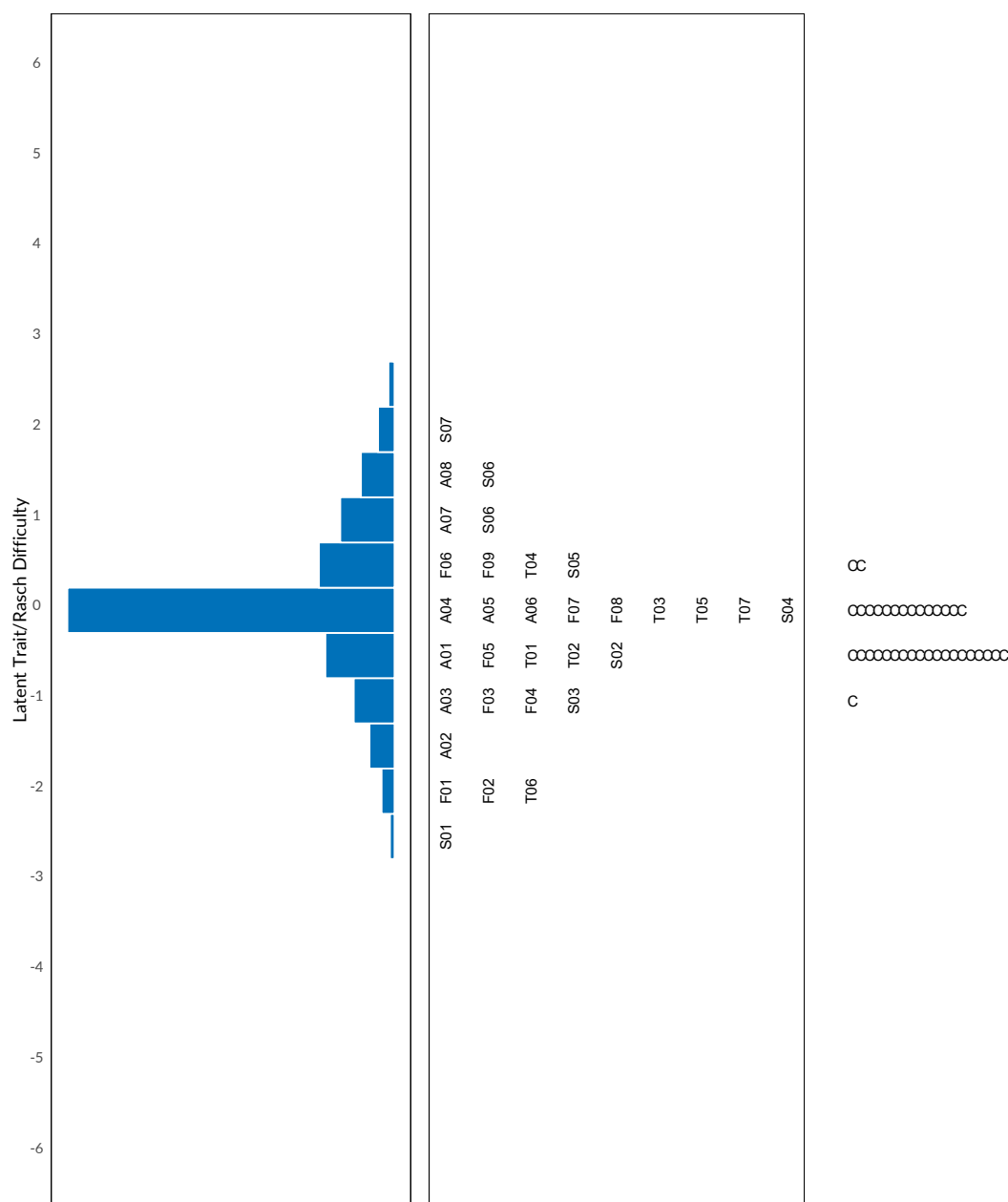




Figure 11.3: Mapping of CT student abilities, item difficulties, and country means



Dimensionality and test reliability

CIL

In ICILS 2023, three new CIL test modules were added to the CIL test instrument. These new modules followed the general structure and integrated the assessment content in ways that were very similar to the CIL test modules from previous cycles. The overall test reliability for CIL following the removal of items with non-satisfactory psychometric properties, based on the pooled dataset and obtained from the scaling model, was 0.86. Given the stability of the test content over time, the similarity of the new test content to existing test content, and the high test reliability, we did not complete further tests of dimensionality for the CIL assessment.

CT

Two new modules were added to the CT assessment. In ICILS 2018, one CT module focused on conceptualization of problems and the second module focused on operationalizing solutions. These two modules were used again in ICILS 2023. The two new CT modules developed for ICILS 2023 integrated tasks that address both conceptualization and operationalization, and included a stronger emphasis on evaluating solutions as part of the iterative problem-solving process.

The differences between the integration of conceptualization and operationalization tasks in the two new CT modules and their greater separation in the two trend CT modules prompted us to carefully evaluate potential dimensionality in the CT data. We fitted a range of different two-dimensional models to the data.

Firstly, we checked each module (dimension one) against the rest (dimension two). Secondly, we split the items based on their categorization, either they contain visual coding (dimension one) or not (dimension two). Finally, we also tested the unidimensional model.

The results suggested that students' CT was well described by using a single dimension given that latent correlations between potential sub-dimensions were very high. The reliabilities based on two dimensional models were mostly lower (between 0.51 and 0.65, with the exception of one module where the reliability was 0.79) than for the one-dimensional model (0.76). With the addition of new modules, the assumption of unidimensionality for CT was revisited and re-examined, with results confirming that the unidimensionality assumption remains valid. The overall test reliability for CT following the removal of items with non-satisfactory psychometric properties, based on the pooled dataset and obtained from the scaling model, was 0.83.

Local dependence

The tasks within each CIL and CT test module are thematically related. In addition, in the ICILS CIL modules, the large tasks are evaluated using sets of analytic scoring criteria. These criteria focus on either the substantive content, or the use of software features associated with the presentation of the information products students produce. Across all cycles of ICILS we have therefore been conscious of the potential for local dependence to affect the data across tasks within each module, and among assessment criteria within each large task. We have consequently checked for any evidence for local dependence within modules and among the large task criteria (Gebhardt & Schulz, 2015; Ockwell et al., 2020).

In ICILS 2023, an analysis of local dependence was conducted by inspecting residual correlations between pairs of items within trend and new modules.

- If justified by the item content, a new scoring rule was developed and used to combine the responses from both items into a single item.
- Alternatively, if necessary, one of the two items was removed from the scale to maintain measurement integrity.

Treatment of item-level non response

In ICILS, students responded to tasks, and the scored data from these tasks (sometimes combined or recoded before scoring) are stored as items for inclusion in the scaling. In this chapter, the term item refers to the item-level data included in the scaling process. Non-response occurs when a student does not provide a response to a task. This can happen for several reasons: the task was not administered by design, the task was administered but the student did not provide an answer, or the task was intended to be administered but was not reached by the student. Tasks contributing to a scored item generally share the same non-response code.

Item response categories included several types of non-response and item score categories. There are three possible types of missing responses:

- “not-administered”: this is used when the student was not administered the task by design, given the matrix-sampling design used by ICILS, where each student is administered only a sample of the assessment tasks (two CIL modules and two CT modules). Additionally, tasks may be recoded as “not-administered” at the international and national levels due to technical failures, incorrect translations, or scaling issues, as previously described.
- “omitted”: this code is used when a student did not respond to an item, based on one or more associated tasks, but did respond to one or more subsequent items within the same module.
- “not-reached”: this code is used when a student did not respond to a item and the item immediately preceding it, and no subsequent items in the module were completed. In such cases, the first unanswered item after the last completed response is coded as “omitted,” and all remaining unanswered items in the module are coded as “not-reached.”

Details of the codes used in the ICILS international database (IDB) to represent item-level non responses are provided in the [ICILS 2023 user guide](#).

In ICILS 2023, items corresponding to tasks that were not reached by students were treated differently when estimating item parameters than when generating student proficiency scores. For item parameter estimation, not-reached responses were ignored and treated as not administered. This approach avoids confounding item difficulty with speededness. In contrast, not-reached items were treated as incorrect when generating student proficiency scores. Omitted items were consistently treated as incorrect during both the calibration and scaling stages.

Item- and task-level non response

CIL

[Table 11.1](#) presents the international percentages of omitted and not-reached CIL items. These are grouped by test module within the table. The modules can be differentiated by the first letter of the item code and the numeral within each item code represents the ordinal position of the related task within the module. Each CIL module contains a large task that is assessed using more than one criteria.²² Non-response codes can either be allocated to all items corresponding to a large task (indicating that the student did not respond to the task) or to none of the items corresponding to the large task (where the student did respond to the task). As a consequence, in [Table 11.1](#) for each large task, all corresponding items show equal percentages of omitted responses. Similarly the items corresponding to each large task show equal percentages of not-reached.

[Table 11.2](#) presents the average percentages of non-response values by item type and for the items within each module for CIL. The omitted and not reached percentages were highest for the large tasks. This is to be expected both because the large tasks are open response format (which typically have higher non-response rates in tests when they are mixed with closed-format items) and also because the large tasks always appear at the end the modules (non-response rates typically tend to increase with ordinal item position in any given test). On average, about eight percent of students omitted an item, while roughly one percent did not reach an item.

The large tasks have between seven and nine corresponding items which may result in inaccurate representations of the extent of task-level non-response within modules. Therefore, [Table 11.3](#) has been included to show the percentages of omitted and not reached task per module.

When comparing omitted and not-reached items and tasks across CIL modules, the “Paper books vs e-books” module had the highest percentage of omitted responses, while the “Internet safety” module had the lowest. The “Recycling” module, which also had the largest number of tasks, showed slightly higher percentages of not-reached responses compared to the other modules.

²² In [Table 11.1](#) the items associated with the large task in each module have the same numeral in their item codes because each item refers to the same task.



CT

[Table 11.4](#) presents the average percentages of missing values by item type and for the items within each CT module. Overall, these percentages were slightly higher than for CIL. Omissions were below 10 percent for approximately two-thirds of the items. In three of the remaining items omissions were above 20 percent. Two of these items were constructed response and the third was an abstraction item (including a set of data recording and comparing actions).

[Table 11.5](#) presents the average percentages of omitted and not reached responses by CT module. Omitted response averages ranged from approximately eight percent to 12 percent. Not reached responses varied from approximately two percent to five percent. As each CT task corresponded to a single item, we have not separately reported omitted and not reached percentages for tasks by module (these data are the same as reported in the bottom section of [Table 11.5](#)).



Table 11.1: Percentages of omitted responses and items not reached for CIL items

Item No.	Item	Omitted	Not Reached	Item Type	Scoring Method
1	C02Z	7.2	0.0	Non-linear skill	Machine
2	C03Z	1.5	0.0	Linear skill	Machine
3	C04Z	1.8	0.3	Non-linear skill	Machine
4	C05Z	2.9	0.7	Non-linear skill	Machine
5	C06Z	10.0	1.5	Constructed response	Human
6	C07A	12.0	5.1	Large task	Human
7	C07B	12.0	5.1	Large task	Human
8	C07C	12.0	5.1	Large task	Human
9	C07D	12.0	5.1	Large task	Human
10	C07E	12.0	5.1	Large task	Human
11	C07F	12.0	5.1	Large task	Human
12	C07G	12.0	5.1	Large task	Human
13	C07H	12.0	5.1	Large task	Human
14	G01Z	5.0	0.0	Constructed response	Human
15	G02Z	4.4	0.0	Linear skill	Machine
16	G03Z	6.6	0.1	Constructed response	Human
17	G04Z	1.8	0.3	MCQ	Machine
18	G05Z	3.5	0.3	Linear skill	Machine
19	G06Z	10.6	0.5	Constructed response	Human
20	G07Z	0.8	0.8	MCQ	Machine
21	G08Z	2.5	0.8	Constructed response	Human
22	G09C	29.1	2.4	Large task	Human
23	G09D	29.1	2.4	Large task	Human
24	G09E	29.1	2.4	Large task	Human
25	G09F	29.1	2.4	Large task	Human
26	G09G	29.1	2.4	Large task	Human
27	H01Z	3.0	0.0	Linear skill	Machine
28	H02Z	1.2	0.0	Linear skill	Machine
29	H03Z	3.4	0.1	Linear skill	Machine
30	H05Z	5.0	1.4	Constructed response	Human
31	H06Z	3.3	2.1	Constructed response	Human
32	H07A	17.0	2.8	Large task	Human
33	H07B	17.0	2.8	Large task	Human
34	H07C	17.0	2.8	Large task	Human
35	H07D	17.0	2.8	Large task	Human
36	H07E	17.0	2.8	Large task	Human
37	H07F	17.0	2.8	Large task	Human
38	H07G	17.0	2.8	Large task	Human
39	H07H	17.0	2.8	Large task	Human
40	H07I	17.0	2.8	Large task	Human



Table 11.1: Percentages of omitted responses and items not reached for CIL items (cont'd)

Item No.	Item	Omitted	Not Reached	Item Type	Scoring Method
41	H07J	17.0	2.8	Large task	Human
42	I01Z	4.0	0.0	Linear skill	Machine
43	I02Z	2.3	0.0	MCQ	Machine
44	I03Z	1.4	0.2	MCQ	Machine
45	I04Z	2.5	0.2	Constructed response	Human
46	I05Z	3.9	0.4	Constructed response	Human
47	I06Z	4.6	1.6	Constructed response	Human
48	I07AB	10.3	3.1	Large task	Human
49	I07C	10.3	3.1	Large task	Human
50	I07D	10.3	3.1	Large task	Human
51	I07E	10.3	3.1	Large task	Human
52	I07F	10.3	3.1	Large task	Human
53	I07G	10.3	3.1	Large task	Human
54	I07H	10.3	3.1	Large task	Human
55	I07I	10.3	3.1	Large task	Human
56	P01Z	20.3	0.0	Linear skill	Machine
57	P02Z	14.2	0.0	Constructed response	Human
58	P03Z	1.8	0.2	MCQ	Machine
59	P04Z	9.3	0.3	Non-linear skill	Machine
60	P05Z	2.9	0.9	Constructed response	Human
61	P06A	17.3	1.9	Large task	Human
62	P06B	17.3	1.9	Large task	Human
63	P06C	17.3	1.9	Large task	Human
64	P06D	17.3	1.9	Large task	Human
65	P06E	17.3	1.9	Large task	Human
66	P06F	17.3	1.9	Large task	Human
67	P06G	17.3	1.9	Large task	Human
68	P06H	17.3	1.9	Large task	Human
69	R01Z	2.5	0.0	MCQ	Machine
70	R02Z	10.5	0.0	Constructed response	Human
71	R03Z	8.5	0.1	Linear skill	Machine
72	R04Z	7.7	0.2	Linear skill	Machine
73	R05A	9.0	0.2	Non-linear skill	Machine
74	R06A	9.7	0.5	Constructed response	Human
75	R06B	9.7	0.5	Constructed response	Human
76	R07Z	9.5	0.9	Non-linear skill	Machine
77	R08Z	12.0	1.1	Constructed response	Human
78	R09Z	2.8	1.6	MCQ	Machine
79	R10Z	2.4	1.8	Constructed response	Human
80	R11A	17.2	2.4	Constructed response	Human

Table 11.1: Percentages of omitted responses and items not reached for CIL items (cont'd)

Item No.	Item	Omitted	Not Reached	Item Type	Scoring Method
81	R11B	17.2	2.4	Constructed response	Human
82	R11C	17.2	2.4	Constructed response	Human
83	R11D	17.2	2.4	Constructed response	Human
84	R12A	10.8	10.3	Large task	Human
85	R12B	10.8	10.3	Large task	Human
86	R12C	10.8	10.3	Large task	Human
87	R12D	10.8	10.3	Large task	Human
88	R12G	10.8	10.3	Large task	Human
89	R12H	10.8	10.3	Large task	Human
90	R12I	10.8	10.3	Large task	Human
91	R12J	10.8	10.3	Large task	Human
92	S01Z	6.7	0.0	Linear skill	Machine
93	S02Z	3.2	0.0	Linear skill	Machine
94	S03Z	28.1	0.1	Linear skill	Machine
95	S04A	2.6	0.2	Non-linear skill	Machine
96	S04B	2.6	0.2	Non-linear skill	Machine
97	S06Z	2.4	0.7	MCQ	Machine
98	S08A	7.5	3.6	Large task	Human
99	S08B	7.5	3.6	Large task	Human
100	S08C	7.5	3.6	Large task	Human
101	S08D	7.5	3.6	Large task	Human
102	S08E	7.5	3.6	Large task	Human
103	S08F	7.5	3.6	Large task	Human
104	S08G	7.5	3.6	Large task	Human



Table 11.2: Percentages of omitted responses and items not reached for CIL items, by item type and module

	Number of items	Omitted	Not Reached
Item Type:			
Constructed response (Human)	21	8.8	1.1
Large task (Human)	54	14.3	4.2
Linear skill (Machine)	13	7.3	0.1
Multiple-choice question (Machine)	8	2.0	0.5
Non-linear skill (Machine)	8	5.6	0.3
Module:			
Board games	13	13.9	1.1
Breathing	15	12.4	2.1
Computer use and health	13	9.2	3.3
Internet safety	14	7.2	1.9
Paper books vs e-books	13	14.4	1.3
Recycling	23	10.4	4.3
School trip	13	7.5	2.0

Table 11.3: Percentages of omitted responses and items not reached for CIL items by module

	Number of tasks	Omitted	Not Reached
Module:			
Board games	9	7.1	0.6
Breathing	6	5.5	1.1
Computer use and health	6	5.9	1.3
Internet safety	7	4.1	0.8
Paper books vs e-books	6	11.0	0.5
Recycling	12	8.6	1.6
School trip	6	8.4	0.8



Table 11.4: Percentages of omitted responses and items not reached for CT items

Item No.	Item	Omitted	Not Reached	Item Type	Scoring Method
1	TA01Z	10.5	0.0	Abstraction	Machine
2	TA02Z	9.9	0.0	Abstraction	Machine
3	TA03Z	25.9	0.5	Abstraction	Machine
4	TA04Z	4.9	1.3	MCQ	Machine
5	TA05Z	7.9	1.5	Drag and drop	Machine
6	TA06Z	2.4	2.6	MCQ	Machine
7	TA07Z	23.0	4.3	Constructed response	Human
8	TA08Z	7.6	19.2	Constructed response	Human
9	TF01E	2.2	0.0	Visual coding	Machine
10	TF02L	3.8	0.0	Visual coding	Machine
11	TF03TEC	3.8	0.6	Visual coding	Machine
12	TF04TEC	4.0	1.1	Visual coding	Machine
13	TF05TEC	5.3	1.9	Visual coding	Machine
14	TF06TEC	10.1	3.1	Visual coding	Machine
15	TF07TEC	12.3	6.5	Visual coding	Machine
16	TF08TEC	17.0	11.0	Visual coding	Machine
17	TF09T	15.3	20.8	Visual coding	Machine
18	TS01Z	2.4	0.0	Drag and drop	Machine
19	TS02Z	2.3	0.0	Drag and drop	Machine
20	TS03Z	8.7	0.4	Numeric entry	Machine
21	TS04Z	5.4	1.5	Visual coding	Machine
22	TS05Z	10.6	3.0	Visual coding	Machine
23	TS06Z	14.7	8.0	Visual coding	Machine
24	TS07Z	30.4	17.2	Constructed response	Human
25	TT01Z	3.3	0.0	Drag and drop	Machine
26	TT02Z	2.7	0.0	MCQ	Machine
27	TT03Z	8.6	0.3	Visual coding	Machine
28	TT04Z	10.6	1.2	Visual coding	Machine
29	TT05B	17.6	2.2	Visual coding	Machine
30	TT06Z	4.4	3.5	Drag and drop	Machine
31	TT07Z	9.1	4.9	Visual coding	Machine



Table 11.5: Percentages of omitted responses and items not reached for CT items, by item type and module

	Number of items	Omitted	Not Reached
Item Type:			
Abstraction (Machine)	3	15.4	0.2
Constructed response (Human)	3	20.3	13.6
Drag and drop (Machine)	5	4.0	1.0
MCQ (Machine)	3	3.3	1.3
Numeric entry (Machine)	1	8.7	0.4
Visual coding (Machine)	16	9.4	4.1
Module:			
Activity tracker	7	8.0	1.7
Automated bus	8	11.5	3.7
Farm drone	9	8.2	5.0
Suns and moons	7	10.6	4.3

Item statistic summaries

To evaluate the psychometric quality of ICILS 2023 achievement items, we calculated a range of country level statistics for each item. These statistics informed the degree to which the data for each item were consistent with the assumptions of the Rasch IRT model described at the beginning of this chapter.

Item statistics were computed for both pooled international data and separately for each of the participating countries. Examples of summary visual displays provided to countries and used by the ISC to support item adjudication are shown for item R02Z in [Figure 11.4](#) and [Figure 11.5](#). The items statistics were based on ICILS 2023 data and include preliminary Rasch analysis statistics, such as Rasch item difficulty, the standard error of this estimate, the Rasch goodness-of-fit index (i.e., infit), gender DIF, and item-by-country interaction.

For human-scored items, the display presents scorer agreement, along with the observed item characteristic curve by category and gender DIF. The second part provides statistics calculated separately for each country.

To assist the reviewing process, the item analysis summary includes a series of ‘flags’ to highlight conditions that may suggest an issue with an item. While these flags do not always indicate a problem, they help identify potential areas of concern, where a consultation with NRCs might be needed. The following conditions were flagged:

- For human-scored items, within-country scoring reliability data showed an agreement of less than 70 percent.
- National item difficulties deviated significantly (by more than 1.3 logits) from the international item difficulty.
- Item discrimination (the point-biserial for the correct answer) was less than 0.2.
- The point-biserial correlation for one or more distractors was greater than zero.
- The point-biserial correlation for the correct response was less than zero.
- Performance or ability estimate was not ordered by score for items with multiple score levels.



- The Rasch goodness-of-fit index was less than 0.88 or greater than 1.12.

As illustrated in Figure 11.4 and Figure 11.5, for each item, the first section presents statistics for the pooled international data, which include the number of student responses, the percentage of students selecting each option or score level, including the percentages of students who omitted or did not reach the item, the percentage of students answering correctly, and the discrimination index (the point-biserial correlation between success on the item and the total score).

Figure 11.4: Item statistics for item R02Z

162

ICILS2023MS-CIL

Recycling (C2R02ZC): Sponsored video label, Constructed response

0=Incorrect; 1=Refers to avoiding advertising; 2=Refers to reliability, commercial intent or relevance

Table 2.12.226: Distribution, point biserial correlation, and average ability by response category for pooled international data

Item	Response category	Score	Students (n)	Percent (%)	Point Biserial Correlation (PB)	Ability
C2R02ZC	0	0	23041	65	-.19	-.35
C2R02ZC	1	1	734	2	.05	.19
C2R02ZC	2	1	7727	22	.35	.54
C2R02ZC	OM	0	3491	11	-.21	-.92

Table 2.12.227: Scorer agreement (available for human scored response), percent correct, Rasch difficulty (standard error), item-rest correlation, infit, and gender DIF for pooled international data

Scorer Agreement	Percent Correct	Rasch Difficulty	(Std. Error)	IRR	Infit	Gender DIF
88	24	.70	(.04)	.32	1.03	.09

Figure 2.12.621: Observed item characteristic curve by category and gender DIF plot

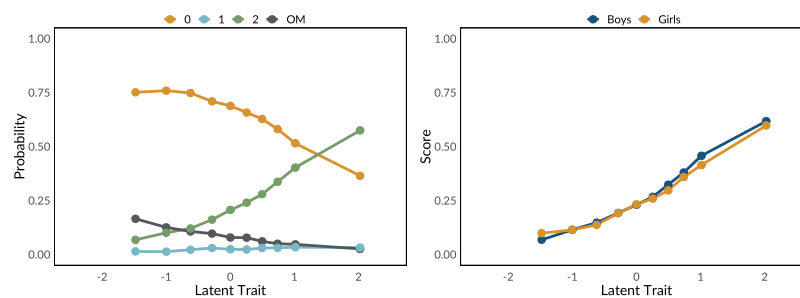


Table 2.12.228: Number of countries that exhibit non-optimal item performance for each criteria

Item	Scorer Agreement <70%	Item-by-Country Interaction			Discrimination				Misfit		
		Country DIF > 1.3	Easier than Intl RDIFF	Harder than Intl RDIFF	Low IRR	Non-key PB is positive	Key PB is negative	Ability not ordered	Small infit (steep slope)	Large infit (flat slope)	RMSE > .12
C2R02ZC	2	0	0	0	2	1	0	0	0	2	1

Intl RDIFF = international Rasch difficulty; IRR = item-rest correlation; PB = point biserial correlation.



Figure 11.5: Item statistics for item R02Z (cont'd)

163

Figure 2.12.631: Percentages of omitted responses and not reached items by country

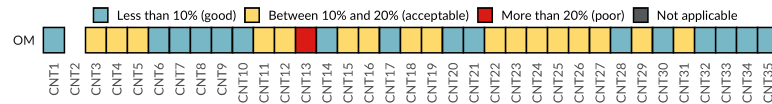


Figure 2.12.632: Country percentage of scorer agreement for human scored response



Figure 2.12.633: Country percent correct

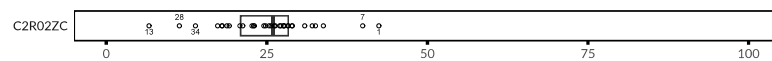


Figure 2.12.634: Item by country interaction (Rasch difficulty)

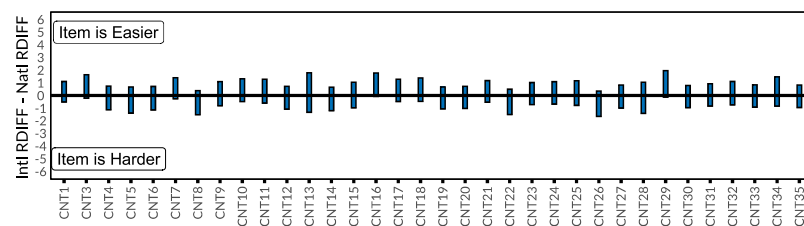


Figure 2.12.635: Country DIF (Intl RDIF - Natl RDIF)

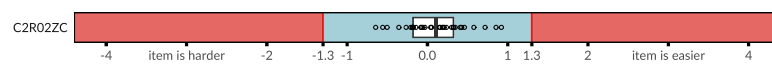


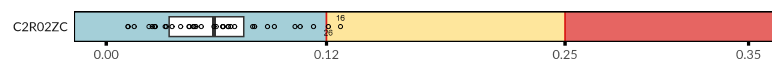
Figure 2.12.636: Country discrimination (item-rest correlation)



Figure 2.12.637: Country item fit (infit)



Figure 2.12.638: Country RMSD



Item fit

The Rasch model assumes that all items are equally discriminating by assuming a slope parameter of 1 so that relatively higher or less discriminating items than the average discrimination across the test show a lack of item fit. Item discrimination refers to the way in which student ability relates to their likelihood of giving correct responses to any given test item. Overall, for any given test item, students of higher ability should be more likely to respond correctly than students of lower ability. In highly discriminating items, the shift between low probability of correct response and high probability



of correct response, takes place over a small range of student abilities. In poorly discriminating items, this shift of probability takes place over a broad range of student abilities. Very low item discrimination occurs when student ability is not or weakly associated with the probability of a correct response on an item. Clearly, low discrimination is not a desirable property for a test item.

In order to judge the measurement properties of test items, it is important to examine the level of discrimination of the items, to flag those for which discrimination is low (because they do not effectively measure the latent trait of CIL or CT), and to consider how well the overall assumption of uniform discrimination is being met by the remaining set of items. Three indicators were used to identify items that violate the equal discrimination assumption: the point-biserial correlation, the weighted mean square statistic (infit), and ICC. The point-biserial correlation is used to measure the strength and direction of the association that exists between a continuous and a dichotomous variable. It represents the correlation of each response category with the test score. We calculated the point-biserial correlation coefficient also for distractors in multiple choice items. The weighted mean square statistic should be close to 1 when an item fits the Rasch model; items with fit statistics lower than 1 discriminate more than expected, whereas those with a fit index greater than 1 discriminate less than expected. While the point-biserial correlation coefficient is affected by ceiling and floor effects (the discrimination can be low if almost all students respond correctly or incorrectly), the weighted mean square statistic is independent of the item difficulty.

In addition, we checked the observed ICC by response categories of the items. The ICC shows the probability of students of different abilities selecting each of the different response options for an item. Ideally, the probability of a correct response should increase, and the probability of an incorrect response should decrease as the ability of student increases (to the right on the horizontal axis). In multiple choice items students with lower ability should be more likely to select options other than the key, but as ability increases the probability of students selecting options other than the key should decrease, while the probability of selecting the key should increase.

Measurement equivalence – Gender DIF

DIF occurs when students with the same ability (represented by their scaled achievement scores) have different probabilities of answering an item correctly. For instance, if boys with the same ability as girls have a higher probability of correctly answering a given item, that item shows DIF with regard to gender. This suggests a violation of the model's assumptions, that the probability is exclusively a function of ability and not of any other characteristics of the respondents.

To estimate gender DIF, interaction terms were included in the item response model. For dichotomous items, gender DIF was modeled as:

$$P_i(\theta_n) = \frac{\exp(\theta_n - (\delta_i - \eta_n + \lambda_{ig}))}{1 + \exp(\theta_n - (\delta_i - \eta_n + \lambda_{ig}))}. \quad (11.8)$$

In this equation, θ_n represents the estimated ability of person n , and δ_i is the estimated difficulty of item i . The term λ_{ig} accounts for the gender effect on the item, and η_g represents the overall gender effect. To obtain accurate estimates, both the item-by-gender interaction (λ_{ig}) and the overall gender effect (η_g) were constrained so that their sum equals zero.

For items with more than two categories, gender DIF was modeled using a partial credit model:

$$P_{x_i}(\theta_n) = \frac{\exp\left(\sum_{j=0}^{x_i} (\theta_n - (\delta_i - \eta_n + \lambda_{ig} + \tau_j))\right)}{\sum_{h=0}^{m_i} \exp\left(\sum_{j=0}^h (\theta_n - (\delta_i - \eta_n + \lambda_{ig} + \tau_j))\right)}, \quad x_i = 0, 1, \dots, m_i, \quad (11.9)$$

where θ_n represents the person's ability, δ_i is the item difficulty on the latent continuum, τ_j is the step parameter, λ_{ig} is the item-by-gender interaction effect, and η_g is the overall gender effect.

We used a DIF value of 0.3 (approximately about one third of a standard deviation) as a criterion for identification of gender DIF. No items showed significant DIF that would warrant their removal from the scaling process in either scale.

Measurement equivalence – Item-by-country interaction

As ICILS is a cross-national study it is important that the test items function similarly across participating countries. When considering the difficulty of test items, the overarching assumption is that the relative difficulty of the test items to each other will be consistent across countries. Item-by-country interaction is evident when, within a given country, a given item is relatively either considerably easier or more difficult in comparison to the other test items than it is, on average, across the remaining countries. Item-by-country interaction is consequently an indicator that the probability of students completing a given item correctly in a given country has a different relationship to student ability in that country than it does across the remaining countries. There can be many causes of item-by-country interaction. In some cases it can be explained by a simple error (such as an error in translation that causes an item to become easier or more difficult in a given country), a problem with scoring (if, for example the item is scored more harshly or leniently in a given country than in other countries), a result of differences in student learning across countries (if, for example, the specific content of an item is a specific strength or weakness of students within a given country) or other unknown factors. One of the aims of the rigorous item development and review process (see [Chapter 2](#)) is to minimize the potential for item-by-country interaction to affect items within countries, however, some degree of item-by-country interaction is inevitable in international large scale assessment (ILSA) and this likelihood increases with the number of countries participating in a study.

Items with a large item-by-country interaction are not suitable to be used in international studies.

The magnitude of item-by-country interaction was estimated by scaling the items separately for each country and then comparing the standardized item difficulty parameter for each item to the corresponding standardized item difficulty parameter calibrated from the pooled data (including all countries meeting the sampling participation requirements). In these displays, the difference between the item difficulty for each country and the international average Rasch item difficulty is represented along with a 95 percent confidence interval. This confidence interval includes a Bonferroni correction²³ to account for multiple comparisons across participating countries. The confidence interval limits were computed as follows:

$$\text{Upper Limit} = RDIFF_{ik} - RDIFF_{\bar{k}} + SE(RDIFF_{ik}) \cdot Z_b \quad (11.10)$$

$$\text{Lower Limit} = RDIFF_{ik} - RDIFF_{\bar{k}} - SE(RDIFF_{ik}) \cdot Z_b, \quad (11.11)$$

where $RDIFF_{ik}$ is the Rasch difficulty of item i in country k , $RDIFF_{\bar{k}}$ is the international average Rasch difficulty of item i , $SE(RDIFF_{ik})$ is the standard error of the Rasch difficulty of item i in country k , and Z_b is the 95 percent critical value from the Z distribution, adjusted for multiple comparisons using the Bonferroni correction.

In addition to graphical displays of item-by-country interaction, as in ICILS 2013 and 2018, country DIF was identified when national item difficulties deviated from the international item difficulty by more than 1.3 logits. These items were excluded from calibrations and scaling at the national level. A list of items that were excluded is available in [Appendix F](#).

²³ When multiple statistical tests are performed such as comparing the item difficulty for each country with the international average Rasch item difficulty, the probability of finding a false positive (rejecting a true null hypothesis) increases. The Bonferroni correction helps counteract this by adjusting the significance level (α) downward.



An example of the graphical representation of item-by-country interaction (Rasch difficulty) for item R02Z is shown in [Figure 11.5](#). This figure illustrates how an item's difficulty varies across countries. The y-axis represents the difference between a country's item difficulty and the international average, while a horizontal line at zero serves as a reference, indicating no deviation from the international difficulty level. The bars represent the standardized national item difficulties, each accompanied by a confidence interval that accounts for statistical uncertainty and is adjusted using a Bonferroni correction for multiple comparisons. If a country's confidence interval does not include zero, it indicates that the item is significantly easier or harder in that country than expected.

Scorer agreement

The ICILS constructed-response items and large tasks were human-scored according to the specifications of the ICILS 2023 scoring guides (see [Chapter 8](#)). New scoring guides were developed for items introduced in ICILS 2023, which were refined after the international field trial. Scoring guides for items already used in ICILS 2018 were kept consistent to ensure that scores retained the same meaning across cycles. In each country, approximately 200 student responses to each task were scored twice. This was done by randomly assigning responses to different scorers through a controlled online scoring system (see [Chapter 3](#)). This double-scoring process allowed for calculation of scorer agreement.

In [Table 11.6](#) and [Table 11.7](#) we present the percentages of scorer agreement for CIL and CT items, based on data from 33 countries²⁴ and one benchmarking entity for CIL, and 23 countries and one benchmarking entity for CT items. The tables show the number of student responses that were double-scored, the international average scorer agreement, and the minimum and maximum agreement values across countries. As has been practice in other IEA studies and previous ICILS cycles, only items that reached a certain threshold of agreement were considered to be used in scaling. The average scorer agreement across countries ranged from 79 for item P06E to 97 percent for items R11A, R11C, and S08B among the 77 constructed response CIL items and 85 for item T3A08Z to 90 percent for item TS07Z among the three CT items. The within country percentage of agreement between scorers on double-scored items across countries ranged from 35 to 100 percent for CIL and 71 to 100 percent for CT items.

The scorer agreement was monitored through the AM Marker system (see [Chapter 3](#)), which was accessible to the lead scorer. In some cases, the scorer agreement displayed in the AM Marker interface was inconsistent with the values calculated by the ISC after scoring was completed. For selected items where the ISC-calculated scorer agreement in a given country was below 70 percent, the scorer agreement shown in the AM Marker system was used instead. All scorer agreement statistics were reviewed and confirmed at the national level. After careful review and discussion, items with scorer agreement below 60 percent were excluded from the scaling process when drawing plausible values for the corresponding national data, although these items remain included in the public database. The list of affected items is available in [Appendix F](#).

24 Chile conducted an additional data collection and thus was excluded from the presentation of scorer agreement.



Table 11.6: Scorer agreement percentages for constructed-response CIL test items

Item	Sample size	Average	Min	Max
C01Z	3,573	88	68	100
C06Z	3,830	87	72	100
C07A	5,192	87	48	100
C07B	5,192	86	57	100
C07C	5,192	85	56	100
C07D	5,192	94	61	100
C07E	5,192	91	67	100
C07F	5,192	87	57	100
C07G	5,192	82	51	100
C07H	5,192	92	63	100
G01Z	4,414	91	74	100
G03Z	5,442	92	77	100
G06Z	4,406	87	68	100
G08Z	4,795	85	56	100
G09C	5,185	84	67	100
G09D	5,185	92	73	100
G09E	5,185	93	81	100
G09F	5,185	94	80	100
G09G	5,185	90	71	100
H05Z	4,062	92	78	100
H06Z	4,463	95	83	100
H07A	5,710	94	75	100
H07B	5,710	82	62	100
H07C	5,710	80	58	100
H07D	5,710	85	62	100
H07E	5,710	87	64	100
H07F	5,710	83	64	100
H07G	5,710	85	51	100
H07H	5,710	88	72	100
H07I	5,710	90	58	100
H07J	5,710	88	58	100
I04Z	4,490	87	69	100
I05Z	5,925	85	64	100
I06Z	5,898	83	56	100
I07A	5,336	87	70	100
I07B	5,336	90	67	100
I07CL*	5,336	93	81	100
I07D	5,336	90	76	100
I07E	5,336	86	63	100
I07F	5,336	85	50	100
I07G	5,336	83	54	100

Note: *Calculated as the mean of the relevant subitem scores.



Table 11.6: Scorer agreement percentages for constructed-response CIL test items (cont'd)

Item	Sample size	Average	Min	Max
I07H	5,336	79	52	100
I07I	5,336	89	57	100
P02Z	4,538	85	49	100
P05Z	4,038	88	68	100
P06AL*	5,295	94	60	100
P06BL*	5,295	94	65	100
P06C	5,295	84	49	100
P06D	5,295	86	60	100
P06E	5,295	79	49	100
P06F	5,295	83	55	100
P06G	5,295	80	47	100
P06H	5,295	85	55	100
R02Z	4,624	88	71	100
R06A	5,453	88	73	100
R06B	5,453	93	71	100
R08Z	4,831	90	80	100
R10Z	4,974	88	68	100
R11A	5,849	97	87	100
R11B	5,849	95	82	100
R11C	5,849	97	86	100
R11D	5,849	96	85	100
R12A	5,278	92	49	100
R12B	5,278	85	56	100
R12C	5,278	84	48	100
R12D	5,278	92	46	100
R12G	5,278	93	56	100
R12H	5,278	86	48	100
R12I	5,278	90	54	100
R12J	5,278	89	35	100
S08A	5,092	91	71	100
S08B	5,092	97	83	100
S08C	5,092	94	76	100
S08D	5,092	93	66	100
S08E	5,092	91	69	100
S08F	5,092	84	64	100
S08G	5,092	85	64	100

Note: *Calculated as the mean of the relevant subitem scores.

Table 11.7: Scorer agreement percentages for constructed-response CT test items

Item	Sample size	Average	Min	Max
T3A07Z	2,406	87	71	100
T3A08Z	2,578	85	72	100
TS07Z	1,727	90	72	100

Test and item evaluation outcomes

Item statistics were reviewed based on the described criteria.

In total, four CIL items (H04A, S05Z, S07Z, C01Z) showed unsatisfactory item statistics, leading to their exclusion from further analysis. Items H04A, S05Z, and S07Z had already been excluded from scaling in 2018 at different stages for calibration and scoring but were readministered in 2023. Item C01Z was newly developed for ICILS 2023. Additionally, due to local dependence, two CIL items, I07A and I07B, were combined into I07AB.

No CT items were removed from scaling in the international pooled data. Items TS06E and TS06F were combined into TS06Z.

[Appendix G](#) provides details about items that were excluded from scaling or removed from the database.

11.4 Implementation of scaling procedures: item calibration

After the test and item adjudication are completed, the assessments are linked for trend scaling. This involves estimating the item parameters for the CIL and CT items in the current assessment by jointly calibrating data from both the current and previous assessments. In 2023, the ICILS concurrent calibration combined achievement data from the 2018 and 2023 assessments. Similar procedures are used in TIMSS, PIRLS, and ICCS (Foy & Yin, 2016, 2017; Ockwell et al., 2020).

The ICILS 2023 data included 30 countries for the CIL assessment, and 21 for the CT assessment, excluding one benchmarking participant and countries that did not meet sampling requirements. The ICILS 2018 calibration sample included 11 countries for CIL (excluding two benchmarking entities and the United States, which did not meet sampling requirements), with seven of these countries also participating in CT. Consequently, concurrent calibration for CIL included student responses from 30 countries in 2023 and 11 from 2018, while the CT calibration incorporated data from 21 countries in 2023 and seven in 2018. Countries were equally weighted within each ICILS cycle for the CIL and CT calibration and all items were included (except for items that were removed nationally or internationally following the adjudication process).

Calibrating the test items

More than half of the CIL items and over one-third of the CT items from the 2018 assessment were reused in the 2023 cycle. To enhance the information available for proficiency estimation and to reduce measurement error, all items from both cycles were included in the calibration process.

Before conducting the concurrent calibration, a separate analysis was carried out to evaluate the quality of the link between the two cycles. Separate calibrations were conducted using all items administered in each respective cycle. Item statistics for the common items were then extracted, reviewed, and compared between 2018 and 2023, both at the international level and separately for each trend country. A total of 64 CIL items and nine CT items were common between the two cycles. This comparison included 10 countries for CIL, Germany, Denmark, Finland, France, Italy, Kazakhstan, Korea (Rep. of), Luxembourg, Portugal, and Uruguay, and seven countries for CT, with Italy, Kazakhstan, and Uruguay opting not to participate in CT.



The left panel of [Figure 11.6](#) presents the item difficulties for CIL common items across the 2018 and 2023 assessment cycles, while the right panel displays the corresponding item difficulties for CT common items. By comparing item difficulties across cycles, this figure provides insights into the stability of item parameters over time and the potential impact of contextual or cohort differences on item functioning. After careful review of the analysis results, eight CIL items (S04B, G03Z, G06Z, G07Z, G09G, R07Z, R10Z, R12J) were retained in the joint calibration dataset but treated as separate items across cycles. Consequently, in the CIL domain, a total of 63 common items were included in the calibration process, of which 55 were used as link items to ensure comparability across cycles, while the remaining eight items were treated differently due to variations in their administration or response patterns.

In CT, all of 11 common items were identified and used as link items to establish score continuity across cycles.

With the linking items defined, item parameters were estimated using multiple-group concurrent calibration based on the Rasch model, with country-by-cycle as the grouping variable. The final sets of CIL and CT item parameters are presented in [Table 11.8](#) and [Table 11.9](#).

Figure 11.6: Item difficulties for common CIL (left) and CT (right) items in 2018 and 2023

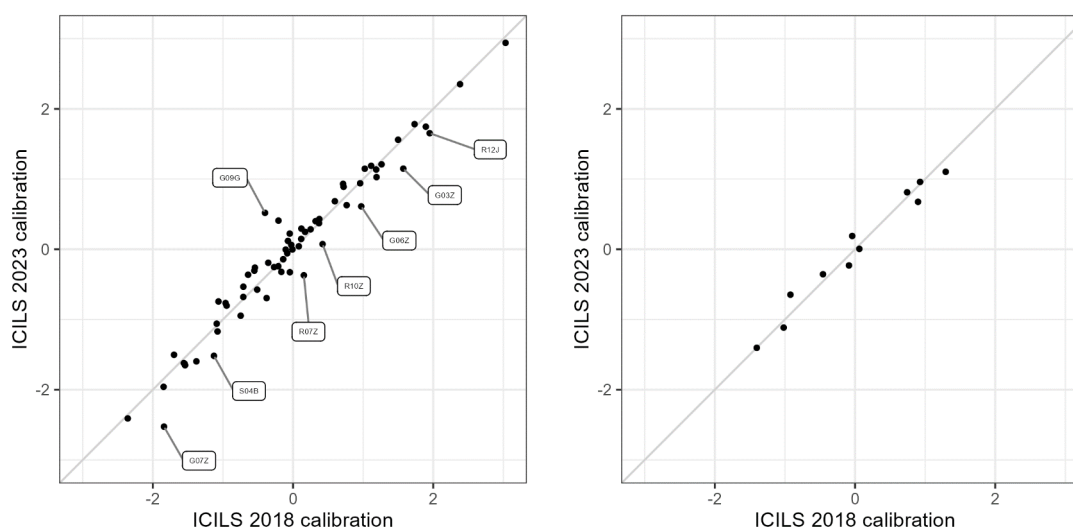




Table 11.8: Item parameter estimates of the CIL test items from concurrent calibration

Item	Item location	Step 1	Step 2	Step 3	Item	Item location	Step 1	Step 2	Step 3
Items common in 2023 and 2018:									
G01Z	2.01				R12G	2.67			
G02Z	0.38				R12H	1.11	0.29	-0.29	
G04Z	-0.63				R12I	2.03			
G08Z	0.77				S01Z	-1.49			
G09C	0.95	-0.25	0.25		S02Z	3.26			
G09D	0.75				S03Z	0.90			
G09E	2.12				S04A	1.77			
G09F	2.40				S06Z	0.64			
H01Z	0.00				S07Z	3.78			
H02Z	1.05				S08A	0.26			
H03Z	-0.95				S08B	0.70			
H05Z	2.14				S08C	0.66	2.70	-2.70	
H06Z	1.15				S08D	0.91	-0.08	0.08	
H07A	-0.69				S08E	1.32	0.89	-0.89	
H07B	0.95	0.10	-0.10		S08F	1.71	-0.46	0.46	
H07C	0.51	-0.14	0.14		S08G	0.59	0.23	-0.23	
H07D	-0.05	0.41	-0.41						
H07E	0.71	0.31	-0.31		Items introduced in 2023:				
H07F	0.98	0.30	-0.30		C02Z	0.10	0.62	-0.62	
H07G	1.83	0.22	-0.22		C03Z	-2.29			
H07H	0.55	0.59	-0.59		C04Z	-0.04	-0.55	0.55	
H07I	0.86				C05Z	-1.92	-0.64	0.64	
H07J	2.12				C06Z	2.30	-0.56	0.56	
R01Z	0.05				C07A	2.83	-0.98	0.98	
R02Z	1.58				C07B	0.74	0.79	-0.79	
R03Z	0.27				C07C	0.39	0.38	-0.38	
R04Z	0.81				C07D	1.01			
R05A	0.34	0.39	-0.79		C07E	0.94			
R06A	-0.12				C07F	1.01			
R06B	1.33				C07G	1.15	0.04	-0.04	
R08Z	0.93				C07H	0.54	1.36	-1.36	
R09Z	-0.21				G03Z	2.06			
R11A	-0.70	2.70	-2.70		G05Z	-0.87	-1.04	1.04	
R11B	-0.67				G06Z	1.50			
R11C	-0.62				G07Z	-1.31			
R11D	0.42				G09G	1.37			
R12A	2.74				I01Z	0.16			
R12B	1.34				I02Z	-2.49			
R12C	0.20				I03Z	0.32			

Table 11.8: Item parameter estimates of the CIL test items from concurrent calibration (cont'd)

Item	Item location	Step 1	Step 2	Step 3	Item	Item location	Step 1	Step 2	Step 3
I04Z	1.54				Items used in 2018 only:				
I05Z	-0.07	-1.24	1.24		B01Z	-0.26			
I06Z	-0.72	-0.31	0.31		B02Z	-0.38			
I07AB	1.00	0.83	-0.97		B03Z	0.87			
I07C	1.63	-0.01	-0.96	0.97	B04Z	1.39			
I07D	2.24	-0.52	0.52		B05Z	0.70			
I07E	1.22				B06Z	-2.06			
I07F	1.53				B07A	-3.16			
I07G	1.33				B07B	-2.71			
I07H	1.32	-0.22	0.22		B07C	-0.24			
I07I	1.66				B08Z	0.79	-0.70	0.70	
P01Z	1.02				B09A	0.85	-1.67	1.67	
P02Z	1.87	-1.39	1.39		B09B	-0.10	1.00	-1.00	
P03Z	0.89				B09C	0.29	0.98	-0.98	
P04Z	1.05	-1.59	0.09	1.49	B09D	0.17			
P05Z	1.91				B09E	1.33			
P06A	1.66	-1.78	-0.19	1.97	B09F	0.84			
P06B	0.67	-0.34	-0.45	0.79	B09G	0.99			
P06C	0.81	1.01	-1.01		G05B	0.17			
P06D	1.63				G03Z_18	2.58			
P06E	1.08	0.38	-0.38		G06Z_18	1.94			
P06F	0.85				G07Z_18	-0.90			
P06G	1.09	-0.14	0.14		G09G_18	0.60			
P06H	0.77	0.67	-0.67		R07Z_18	1.09			
R07Z	0.48				R10Z_18	1.29			
R10Z	0.90				R12J_18	2.90			
R12J	2.51				S04B_18	-0.24			
S04B	-0.76								



Table 11.9: Item parameter estimates of the CT test items from concurrent calibration

Item	Item location	Step 1	Step 2	Step 3
Items common in 2018 and 2023:				
TA01Z	0.10	0.56	-0.56	
TA02Z	-0.73	0.32	-0.32	
TA03Z	-0.31	1.28	-1.28	
TA04Z	0.96			
TF01E	-1.00	-1.32	0.10	1.22
TF02L	-1.32	0.80	-0.80	
TF03TEC	-0.11	-0.39	-1.48	1.86
TF05TEC	0.09	-0.94	0.64	0.29
TF07TEC	0.86	-0.44	0.44	0.00
TF08TEC	0.84	2.18	-2.18	
TF09T	1.22	1.76	-1.76	
Items introduced in 2023:				
TA05Z	0.86	0.58	-0.58	
TA06Z	0.74	0.39	-0.39	
TA07Z	1.69	0.76	-0.76	
TA08Z	1.98	-0.74	0.74	
TF04TEC	-0.42	0.37	-1.46	1.09
TF06TEC	1.25	-1.16	-0.59	1.75
TS01Z	-1.57	-2.02	2.02	
TS02Z	0.06	1.77	-1.77	
TS03Z	-0.45	-0.42	0.42	
TS04Z	0.52			
TS05Z	1.35	3.17	-3.17	
TS06Z	1.54	0.76	-0.46	-0.30
TS07Z	2.51	-0.19	0.19	
TT01Z	0.33	-1.70	1.70	
TT02Z	0.47	1.40	-1.40	
TT03Z	0.59			
TT04Z	1.11	-0.81	0.81	
TT05B	0.61	-0.47	0.47	
TT06Z	-1.16			
TT07Z	0.80	-0.76	0.76	
Items used in 2018 only:				
TA05Z_18	0.48	-0.04	0.04	
TA06Z_18	0.26	0.24	-0.24	
TA07Z_18	1.34	0.72	-0.72	
TA08Z_18	1.48	-0.76	0.76	
TF04TEC_18	-0.32	0.05	-1.50	1.45
TF06TEC_18	0.88	-0.74	-0.59	1.33



11.5 Implementation of scaling procedures: population modeling to estimate students' scores

Principal component analysis

To select and prepare the variables to be used in the latent regression model needed to generate plausible values (PVs), some pre-processing is needed. The pre-processing varied for different variable types:

- Categorical variables with fewer than eight response categories were dummy-coded, including categories representing missing data. For a variable with N categories, N-1 binary variables were created, with one category omitted as the reference to avoid multicollinearity. For example, a variable indicating school type (public versus private) was transformed into a binary indicator, with 1 representing public schools and 0 otherwise. If data were missing, an additional binary variable flagged these cases (1 for missing, 0 otherwise).
- Continuous variables were retained in their original form, with missing values imputed using an appropriate estimate (e.g., student age), and a dummy code was added to indicate those cases with missing values in the original variable.

Given the large number of variables, principal component analysis was employed to reduce the number of variables needed by identifying a smaller number of orthogonal components that account for most of the variation in the context variables. Principal components that explained 90 percent of the variance in all context variables were retained as conditioning variables. To prevent overfitting, the number of retained components was also limited to no more than five percent of a country's student sample size. Therefore, principal components were selected that either retained 90 percent of common variance or five percent of the sample size, whichever led to fewer principal components. Because principal component analysis was conducted separately for each country, the number of components retained varied accordingly.

Additionally, conditioning variables included some direct regressors, specifically student gender, explicit stratum indicators, and the adjusted school mean achievement (the average of weighted likelihood ability estimates (WLEs) for all other students in the same school).

Note that the principal component analysis was conducted separately for each country based on international variables (collected by every participating country) as well as variables with country-specific response categories, e.g., country of birth. [Table 11.10](#) provide details of the conditioning variables used for proficiency estimation. [Appendix E](#) provides a comprehensive list of the student variables used for conditioning, with their respective values.



Table 11.10: Conditioning models for CIL and CT achievement scales

Country	2023				2018			
	Sample size	Number of Principal Components Available	Number of Principal Components Retained	Percentage of Variance Explained	Sample size	Number of Principal Components Available	Number of Principal Components Retained	Percentage of Variance Explained
Austria	3,448	577	172	71.0				
Azerbaijan	3,634	530	181	76.5				
Belgium (Flemish)	3,365	534	168	73.4				
Bosnia and Herzegovina	1,877	531	93	59.8				
Chile	3,216	522	160	75.4	3,092	531	154	70.3
Chinese Taipei	5,112	533	255	85.6				
Croatia	2,911	531	145	73.4				
Cyprus	3,182	531	159	73.8				
Czech Republic	8,169	539	324	90.1				
Denmark	3,038	530	151	71.0	2,404	544	120	59.5
Finland	4,249	534	212	82.8	2,546	547	127	64.7
France	3,694	524	184	75.4	2,940	536	146	66.7
Germany	5,065	543	253	86.5	3,655	559	182	74.4
Greece	3,576	557	178	71.6				
Hungary	3,491	522	174	73.1				
Italy	3,376	520	168	70.5	2,810	534	140	62.0
Kazakhstan	4,852	530	242	82.4	3,371	539	168	71.1
Korea, Republic of	3,723	532	186	79.5	2,875	538	143	66.7
Kosovo	3,345	530	167	70.3				
Latvia	2,705	533	135	68.2				
Luxembourg	4,703	553	235	82.2	5,401	556	270	84.8
Malta	3,115	529	155	74.5				
Netherlands	1,288	530	64	57.0				
North-Rhine W. (Germany)	2,726	543	136	69.0	1,991	557	99	57.9
Norway (Grade 9)	4,436	525	221	83.9				
Oman	6,437	526	319	90.1				
Portugal	3,650	531	182	75.7	3,221	540	161	69.4
Romania	3,270	531	163	75.7				
Serbia	3,125	522	156	73.1				
Slovak Republic	3,034	526	151	68.2				
Slovenia	3,318	533	165	74.5				
Spain	11,799	530	296	90.0				
Sweden	3,401	517	170	77.6				
United States	2,352	524	117	68.0	6,790	532	297	90.0
Uruguay	2,933	531	146	68.1	2,613	541	130	66.0



Latent regression analysis

Latent regression analysis with IRT item parameter estimates ($\mathbf{X}$) treated as fixed values and the principal components of the context questionnaire variables as predictors ($\mathbf{Y}$) for estimating the latent regression parameters $\mathbf{\Gamma}$ (regression coefficients) and $\mathbf{\Sigma}$ (residual variance–covariance matrix). Note that latent regression models are estimated separately for each country to allow for differences in associations between the context characteristics and the cognitive skills.

The TAM package used for analysis supports two-dimensional latent regression models by using item responses across the CIL and CT scales and the correlation between these scales, thereby enhancing the reliability of each scale. For countries administering both CIL and CT, a two-dimensional model was applied; for those administering only CIL, a one-dimensional model was used.

The inputs include students' responses to administered items, item parameters estimated during concurrent calibration, and selected conditioning variables. The output includes estimated regression effects, the residual variance-covariance matrix, and posterior distribution of each student's latent ability (CIL and, where relevant, CT) (see, for example, (Mislevy, 1991; von Davier & Sinharay, 2013; von Davier et al., 2006)). The posterior distribution is a probability distribution of a students' true ability and is based on their responses to the test and questionnaire items.

Drawing plausible values to estimate students' scores

Plausible values are randomly drawn from the resulting posterior distribution for each of the participating students. Additionally, the same procedure were applied to the 2018 data for trend countries using parameters from concurrent calibration, allowing the generation of plausible values known as "ICILS 2018 data from the ICILS 2023 calibration."

When the estimation of the regression parameters ($\mathbf{\Gamma}$ and $\mathbf{\Sigma}$) is complete, plausible values are randomly drawn from the joint distribution of the values of $\mathbf{\Gamma}$ for all sampled respondents in a three-step process:

1. First, a value of $\mathbf{\Gamma}$ is drawn from a normal approximation to $P(\mathbf{\Gamma}, \mathbf{\Sigma} \mid \mathbf{x}_j, \mathbf{y}_j)$ that fixes $\mathbf{\Sigma}$ at the value $\hat{\mathbf{\Sigma}}$.
2. Second, conditional on the generated value of $\mathbf{\Gamma}$ (and the fixed value of $\mathbf{\Sigma} = \hat{\mathbf{\Sigma}}$), the mean and variance of the posterior distribution of $\boldsymbol{\theta}$ are computed using the same methods applied in the EM algorithm.
3. In the third step, the $\boldsymbol{\theta}$ is drawn independently from a multivariate normal distribution with mean and variance.

These three steps are repeated five times, resulting in five independent plausible values (PVs) of $\boldsymbol{\theta}$ given $\mathbf{x}$ and $\mathbf{y}$ for each sampled student in CIL and CT who participated in both assessments, or five PVs of CIL for students who only participated in CIL.

Transforming proficiency scores to measure trends

For comparability with previous ICILS assessments, the 2023 CIL and CT scale scores were transformed to the metric used in 2018 as part of the concurrent calibration approach. Linear transformations constants needed for placing the 2023 scores to the trend metric were based on the subset of countries that met sampling participation requirements and participated in both ICILS cycles, 2018 and 2023. Finally, 10 countries contributed to the CIL transformation constants (Germany, Denmark, Finland, France, Italy, Kazakhstan, Korea (Rep. of), Luxembourg, Portugal, and Uruguay), and seven to the CT ones (Germany, Denmark, Finland, France, Korea (Rep. of), Luxembourg, and Portugal). Linear transformations matched the means and standard deviations of ICILS 2018 data from the ICILS 2023 and 2018 calibrations and ensured alignment for trend reporting. This resulted in directly comparable CIL and CT achievement scores across the 2018 and 2023 assessments.



These linear transformations were given by:

$$PV_{ik}^* = A_k + B_k \times PV_{ik}, \quad (11.12)$$

where

- PV_{ik} is the ICILS 2018 plausible value i of scale k prior to transformation;
- PV_{ik}^* is the ICILS 2018 plausible value i of scale k after transformation; and
- A_k and B_k are the linear transformation constants.

The linear transformation constants were obtained by first computing the international means and standard deviations of the plausible values for the CIL and CT scales using the plausible values produced in 2018 based on the 2018 item calibrations for the trend countries that met sampling requirement, resulting in 10 CIL countries and seven CT countries. These were the plausible values published in 2018.

Next, the same calculations were done using the plausible values from “ICILS 2018 data from the ICILS 2023 calibration” for the same set of trend countries. From these calculations, the linear transformation constants were obtained as:

$$B_k = \frac{\sigma_k}{\sigma_k^*} \quad (11.13)$$

$$A_k = \mu_k - B_k \cdot \mu_k^*, \quad (11.14)$$

where

- μ_k is the international mean of scale k averaged across the five plausible values published in 2018;
- μ_k^* is the international mean of scale k averaged across the five plausible values from the 2018 assessment based on the 2023 concurrent calibration;
- σ_k is the international standard deviation of scale k averaged across the five plausible values i published in 2018;
- σ_k^* is the international standard deviation of scale k averaged across the five plausible values from the 2018 assessment based on the 2023 concurrent calibration.

Table 11.11 shows the international mean and standard deviation of plausible value published in 2018 as well as the plausible values from the 2018 assessment based on the 2023 concurrent calibration. Table 11.12 summarizes the ICILS 2023 transformation constants for both scales.



Table 11.11: ICILS 2018 published scores and re-scaled scores

	CIL				CT			
	ICILS 2018 published results		ICILS 2018 re-scaled results		ICILS 2018 published results		ICILS 2018 re-scaled results	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PV1	494.7391	96.7024	0.2868	1.1389	500.1119	99.8556	0.1111	0.9621
PV2	494.7076	96.9165	0.2890	1.1436	499.9198	100.0711	0.1138	0.9621
PV3	494.6434	97.0379	0.2876	1.1385	500.0882	99.9607	0.1152	0.9589
PV4	494.7649	97.0348	0.2893	1.1423	499.8380	100.1688	0.1150	0.9606
PV5	494.7581	96.7471	0.2925	1.1384	500.0422	99.9320	0.1146	0.9670
Average	494.7226	96.8877	0.2890	1.1403	500.0000	99.9976	0.1140	0.9621

Table 11.12: Transformation constants for the ICILS 2023

	Sample Size	ICILS 2018 published results		ICILS 2018 re-scaled results		Transformation Constants	
		Mean	SD	Mean	SD	A	B
CIL	31,836	494.7226	96.8877	0.2890	1.1403	470.1649	84.9639
CT	23,042	500.0000	99.9976	0.1140	0.9621	488.1557	103.9340

Estimating linking error

The transformation of proficiency scores allows us to place the scores of ICILS 2023 to the trend scale established in earlier ICILS cycles. One way to evaluate the transformation is to calculate the difference between linking item parameters obtained from 2023 with those reported in ICILS 2018. The sample of items chosen as link items might influence the transformation. This means that the resulting transformation would have been slightly different if we had chosen an alternative set of link items. Uncertainty in the transformation thus relates to the sampling of the link items, in the same way that uncertainty in values such as country averages is an outcome of the particular sample of students that is used.

The uncertainty resulting from link-item sampling is referred to as linking error, and it is an error that analysts have to take into account when comparing the results arising out of different data collections (see (Monseur & Bereznier, 2007)). As is the situation with the error that is introduced through the process of sampling students, the exact magnitude of this linking error cannot be determined. We can, however, estimate the likely range of magnitudes for this error and take it into account when interpreting results. As with sampling errors, the likely range of magnitude for the errors is represented as a standard error.

To calculate linking error between the 2023 and 2018 CIL and CT, we first estimated the Rasch difficulty parameters for the 2023 assessments. The common item parameters were then centered by setting the mean to 0. The Rasch difficulty parameters for the common items in the 2018 administration were also centered by setting the mean to 0. The difference between the two item parameters was calculated and averaged across all items, thus representing the difference in relative difficulty of the common items across cycles. Finally, linking error was computed as the standard error of the mean difference using the traditional formula for calculating the standard error of a mean. This approach assumes that the common items are a random sample of all possible common items.

It is evident that linking error in ICILS is a function of item drift (differences in item difficulty from one administration to the next), the clustering of items within units, the weighting of maximum points for polytomous items, and the number of common items across cycles.

Suppose we have a total of L score points in the link items in K modules. Use i to index items in a unit and j to index units so that δ_{ij}^y is the estimated difficulty of item i in unit j for year y , and let:



$$\hat{\delta}_{ij}^{2023} - \hat{\delta}_{ij}^{2018}. \quad (11.15)$$

The size (number of score points) of unit j is m_j so that:

$$\sum_{j=1}^K m_j = L \quad \text{and} \quad \bar{m} = \frac{1}{K} \sum_{j=1}^K m_j. \quad (11.16)$$

Further let:

$$c_j = \frac{1}{m_j} \sum_{i=1}^{m_j} c_{ij}, \quad \text{and} \quad \bar{c} = \frac{1}{N} \sum_{j=1}^K \sum_{i=1}^{m_j} c_{ij}, \quad (11.17)$$

and then the link error, taking into account the clustering of items was computed as follows:

$$\text{LinkError}_{2023,2018} = \sqrt{\frac{\sum_{j=1}^K m_j^2 (c_j - \bar{c})^2}{K(K-1)\bar{m}^2}} = \frac{\sum_{j=1}^K m_j^2 (c_j - \bar{c})^2}{L^2} \frac{K}{K-1}. \quad (11.18)$$

Therefore, when comparing scale scores between 2023 and 2018, the equating error is 2.8394 scale points for the CIL scale and 2.5686 scale points for the CT scale.

Table 11.13 and Table 11.14 present the differences between countries' published and re-estimated 2018 results. These results compare the published 2018 scores with the re-estimated 2018 scores based on the concurrent calibration conducted in 2023, alongside the ICILS 2023 results. Overall, there was good agreement between the published and re-estimated results. Although small differences were observed at the country level, most differences in CIL were within one point, and no standard errors exceeded 0.6 points. In CT, absolute differences reached up to 3 points, with no standard errors exceeding 1.6 points. These minor differences are expected, given the larger number of participating countries in 2023, the numbers of principal components used for conditioning, and changes in the software used.

Table 11.13: Trend linking error variance between ICILS 2018 and ICILS 2023 calibrations in the CIL achievement scale

Country	ICILS 2018 published results				ICILS 2018 re-scaled results				Difference	
	Average	(s.e)	SD	(s.e.)	Average	(s.e)	SD	(s.e)	Average	(s.e)
Denmark	553	(2.0)	66	(1.4)	554	(2.1)	66	(1.4)	-1	(0.6)
Finland	531	(3.0)	81	(2.4)	532	(2.8)	80	(1.9)	-1	(0.6)
France	499	(2.3)	80	(2.2)	498	(2.4)	82	(1.9)	1	(0.6)
Germany	518	(2.9)	80	(2.9)	518	(2.8)	79	(2.7)	0	(0.5)
Italy	461	(2.8)	82	(1.7)	462	(2.8)	80	(1.7)	-1	(0.4)
Kazakhstan	395	(5.4)	106	(3.2)	395	(5.5)	105	(3.3)	0	(0.5)
Korea, Republic of	542	(3.1)	95	(1.8)	541	(3.1)	94	(1.5)	1	(0.4)
Luxembourg	482	(0.8)	84	(0.8)	482	(0.7)	84	(0.7)	-1	(0.3)
Portugal	516	(2.6)	71	(1.8)	517	(2.9)	72	(2.2)	0	(0.6)
Uruguay	450	(4.3)	101	(2.0)	449	(4.3)	101	(2.5)	1	(0.5)



Table 11.14: Trend linking error variance between ICILS 2018 and ICILS 2023 calibrations in the CT achievement scale

Country	ICILS 2018 published results				ICILS 2018 re-scaled results				Difference	
	Average	(s.e)	SD	(s.e.)	Average	(s.e)	SD	(s.e)	Average	(s.e)
Denmark	527	(2.3)	84	(1.6)	527	(2.5)	86	(1.5)	0	(0.7)
Finland	508	(3.4)	97	(2.1)	506	(3.6)	102	(4.2)	2	(0.8)
France	501	(2.4)	92	(2.0)	503	(2.8)	95	(2.1)	-2	(0.7)
Germany	486	(3.6)	110	(2.3)	486	(3.7)	102	(2.8)	0	(1.1)
Korea, Republic of	536	(1.4)	111	(2.4)	534	(1.5)	105	(1.8)	2	(0.6)
Luxembourg	460	(9.0)	107	(10.4)	463	(8.8)	109	(8.9)	-3	(1.6)
Portugal	482	(2.5)	80	(1.9)	483	(2.4)	77	(1.7)	-1	(0.6)

11.6 Proficiency levels for CIL and CT

One of the objectives of ICILS is to establish clearly defined CIL and CT scales that serve as reference points for future international assessments in this learning area. Defining proficiency levels provides a meaningful way to describe student performance across countries and sets benchmarks for future surveys.

The ICILS reporting scales for CIL and CT were established in 2013 and 2018 respectively. Each of these scales were established with a mean of 500 and a standard deviation of 100 for equally weighted national samples for countries that had met the IEA sample participation requirements (Fraillon et al., 2014, 2020).

The described proficiency levels for CIL were established as part of ICILS 2013. The level boundaries have remained the same since that time. The level descriptions, and examples of student achievement at each level are reviewed and, where appropriate, revised with as part of each ICILS cycle to ensure that the content of the scale reflects the new content included in each new cycle.

In ICILS 2018, we established descriptions of three regions of achievement on the CT scale. Because these descriptions were based on item content collected from two modules and seven countries only, the descriptions were reported as preliminary, and with a view to them being revised and consolidated in ICILS 2023 (Fraillon, 2025). This work was completed as part of ICILS 2023. The CT described proficiency levels are presented and discussed in detail in the ICILS 2023 international report (Fraillon, 2025), and, similar to the CIL described levels, will be reviewed and revised with each new ICILS cycle.

The same processes was used to establish the CIL and CT described levels of proficiency.

Two parameters affecting decisions about the level widths and descriptions were first established.

1. The level widths would be equal. While it is feasible to establish described levels with unequal widths on the scale, this would lead to inconsistent interpretations of what it means for a student to “be at a level.”
2. The response probability (rp) for reporting item parameters was set at $rp = 0.62$. In practice, this means that a student with an estimated ability equal to the difficulty of a given item would have a 62 percent probability of correctly responding to the item. This reflects a higher and more appropriate level of “mastery” than $rp = 0.5$.

With these parameters established, the process of establishing the level widths and descriptions began. This included the following iterative processes:

- Reviewing the content and scaled difficulty of the test items to identify themes, content and concepts that contribute to their difficulty
- Characterizing common difficulty-related characteristics of items within difficulty ranges on the



scale

- Establishing equally-spaced boundaries on the scale and review the characteristics of the items within each “level”
- Continuing to adjust the boundaries until “the content of each level showed distinct characteristics and the differences between each level demonstrated a clear progression from lower to higher achievement” (Fraillon, 2025, p. 94)

While in theory, according to the above decisions and processes the level widths could be very small or very large, one final requirement was used to inform the level width.

This relates to what can be inferred about a student whose score places them within a given level. Ideally, when a student’s score places them within a level, one can infer that the student can correctly respond to a reasonable proportion of test content within that level. Students with scores near the top of a level could be expected to complete a larger proportion of content than students with scores near the bottom. However, if a student with a score near the bottom of a level is to be described as being “within the level” then it is reasonable to assume that this student has mastery of some of the content within that level.

For the purpose of evaluating the viability of level widths, it is possible to estimate the expected success of students with scores at different points within a level of a given width, on a hypothetical test containing items with difficulties distributed evenly across the level.

Ultimately, following iterative processes to determine level content and widths, both the CIL and CT levels were established to have widths of 1.00 logits on the ICILS CIL and CT scales.

Using these parameters, we were able to infer the following about students’ CIL and CT in relation to their respective proficiency levels:²⁵

- Students whose scores placed them at the lowest possible point of a proficiency level were likely to correctly answer (on average) approximately 50 percent of the items (on a hypothetical test made up of items with locations spread uniformly across the level).
- Students whose scores placed them at the top of a proficiency level were likely to correctly answer (on average) approximately 72 percent of the items (on a hypothetical test made up of items with locations spread uniformly across the level).

When reporting released CIL items and mapping them against proficiency levels, we had to transform location parameters of these items to a value that reflected a response probability of 62 percent ($rp = 0.62$). This is achieved by adding the natural log of the odds of a 62 percent chance to the original log odds and transforming the result to the international metric by applying the same transformation as for the (original) student scores. The standardized item difficulty δ'_i for each item was obtained as follows:

$$\delta'_i = \mu_k + \sigma_k \times \left(\frac{(\delta_i + \ln(\frac{0.62}{0.38})) - \mu_k^*}{\sigma_k^*} \right) \quad (11.19)$$

Here, δ_i is the item difficulty in its original metric. The constants μ_k , μ_k^* , σ_k , and σ_k^* are defined in the same way as those used for transforming proficiency scores to measure trends. Specifically, μ_k is the international mean of scale k , averaged across the five plausible values published in 2018. μ_k^* is the international mean of scale k , averaged across the five plausible values from the 2018 assessment based

²⁵ Please note these inferences are only valid with respect to proficiency levels that have both a defined lower and upper boundary.



on the 2023 concurrent calibration. σ_k is the international standard deviation of scale k , averaged across the five plausible values published in 2018. σ_k^* is the international standard deviation of scale k , based on the 2023 concurrent calibration of the 2018 assessment. The constants are documented in [Table 11.11](#) and [Table 11.12](#).

The transformation for each CIL item was obtained as follows:

$$\delta'_i = 494.7226 + 96.8877 \times \left(\frac{(\delta_i + \ln(\frac{0.62}{0.38})) - 0.2890}{1.1400} \right) \quad (11.20)$$

The transformation for each CT item was obtained as follows:

$$\delta'_i = 500.0000 + 99.9976 \times \left(\frac{(\delta_i + \ln(\frac{0.62}{0.38})) - 0.1140}{0.9621} \right) \quad (11.21)$$

For CIL, the final level boundaries established were set at 407, 492, 576, and 661 scale score points on the CIL scale, with a corresponding level width of 85 scale score points for each level with an upper and lower boundary. The levels are: Below Level 1 (407 scale score points and below), Level 1 (above 407 to 492 scale score points), Level 2 (above 492 to 576 scale score points), Level 3 (above 576 to 661 scale score points), and Level 4 (above 661 scale score points).

For CT, the final level boundaries were set at 330, 440, 550, and 660 scale score points on the CT scale, with a corresponding level width of 110 scale score points for each level with an upper and lower boundary. The levels are: Below Level 1 (330 scale score points and below), Level 1 (above 330 to 440 scale score points), Level 2 (above 440 to 550 scale score points), Level 3 (above 550 to 660 scale score points), and Level 4 (above 660 scale score points).



References

- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315814698>
- Foy, P., & Yin, L. (2016). Scaling the TIMSS 2015 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center at Boston College. <https://timssandpirls.bc.edu/publications/timss/2015-methods.html>
- Foy, P., & Yin, L. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center at Boston College. <https://timssandpirls.bc.edu/publications/pirls/2016-methods.html>
- Fraillon, J. (Ed.). (2025). *An international perspective on digital literacy: Results from ICILS 2023*. Springer. <https://doi.org/10.1007/978-3-031-87722-3>
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: IEA International Computer and Information Literacy Study 2018 international report*. Springer Nature. <https://doi.org/10.1007/978-3-030-38781-5>
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA International Computer and Information Literacy Study international report*. Springer Nature. <https://doi.org/10.1007/978-3-319-14222-7>
- Gebhardt, E., & Schulz, W. (2015). Scaling procedures for ICILS test items. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *ICILS 2013 technical report* (pp. 155–174). International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2013-technical-report>
- Linacre, M. D. (2017). Teaching Rasch measurement. *Rasch Measurement Transactions*, 31(2), 1630–1631.
- Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). Springer. https://doi.org/10.1007/978-1-4757-2691-6_6
- Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80(392), 993–997. <https://doi.org/10.1080/01621459.1985.10478215>
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196. <https://doi.org/10.1007/BF02294457>
- Mislevy, R. J., & Sheehan, K. (1987). Marginal estimation procedures. In A. Beaton (Ed.), *The NAEP 1983–1984 technical report* (pp. 293–360). National Assessment of Educational Progress, Educational Testing Service.
- Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, 8(3), 323–335.
- Ockwell, L., Daraganov, A., & Schulz, W. (2020). Scaling procedures for ICILS 2018 test items. In J. Fraillon, J. Ainley, W. Schulz, T. Friedman, & D. Duckworth (Eds.), *ICILS 2018 technical report* (pp. 133–158). International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test analysis modules*. <https://CRAN.R-project.org/package=TAM>
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI monograph series*, 2(1), 9–36.
- von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 155–174). Chapman; Hall/CRC. <https://doi.org/10.1201/b16061>
- von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). *Stochastic approximation methods for latent regression item response models* (Research Report No. RR-06-08). Educational Testing Service.

Chapter 12:

Scaling procedures for ICILS 2023 questionnaire items

Minge Chen, Andrés Christiansen, Mojca Rožman, and Rolf Strietholt

This chapter describes the procedures for the scaling of the ICILS 2023 context questionnaire data (collected from students, teachers, school principals, and ICT coordinators) and the indices that were derived from these data. Three types of indices were computed with questionnaire data for ICILS 2023:

- *Simple indices* were constructed through arithmetical transformation or recoding, for example an index of immigration background based on information about the country of birth of students and their parents;
- *Students' socioeconomic background indices* were constructed via principal component analysis (PCA) of three simple indices;
- *Scale indices* were derived through item response theory (IRT) modeling.

This chapter is divided into four sections. The first section describes the construction of simple indices; the second section explains the procedure for the calculation of the students' socioeconomic background indices; the third section includes the theoretical framework underpinning, and the procedure for scaling and scoring scale indices; and the last section lists the scaling results of questionnaire scales.

Questions from the questionnaires are referred to throughout this chapter. For brevity, the following abbreviations are used to denote the questionnaires: student questionnaire (StQ), teacher questionnaire (TQ), principal questionnaire (PQ) and ICT coordinator questionnaire (CQ). Questions within each questionnaire are referred by their number only, for example, Question 5 in the student questionnaire is referred to as StQ 5, or Question 10 in the teacher questionnaire as TQ 10.

12.1 Simple indices

Student questionnaire

Student age

Student age (S_AGE) was calculated as the difference between the year and month of the testing and the year and month of a student's birth. Information from StQ 1 was used to derive age, except for students where this information was missing. In these cases information from student tracking forms provided data for the calculation of this index.

The formula for computing S_AGE was:

$$S\_AGE = (T_y - S_y) + \frac{(T_m - S_m)}{12},$$

where T_y and S_y are, respectively, the year of the test and the year of birth of the tested student, in four-digit format (e.g., 2023 or 2010), and where T_m and S_m are respectively the month of the test and the month of the student's birth. The result was rounded to two decimal places.

Sex of student

The sex of students (S_SEX) was obtained from StQ 2, where the students were asked about their gender. Girl was coded as 1 while boy was coded as 0. For students with omitted data for this question,



the gender from the tracking form was included. Countries could adapt the responses to include one or more additional categories for students to self-identify as neither male nor female. Few students responded using the additional categories in countries where the national option was implemented, during data processing data in any additional categories were recoded to missing.

Students' expected education

Students' expected education (S_ISCED) was obtained from StQ 3, where the students were asked about their expected highest level of educational attainment. The responses were classified using the International Standard Classification of Education (ISCED) framework (UNESCO Institute for Statistics, 2012) and had the following categories (in parentheses the code for each category in the dataset is presented):

- (0) No completion of ISCED level 2;
- (1) Completion of ISCED level 2 (lower-secondary);
- (2) Completion of ISCED level 3 (upper-secondary);
- (3) ISCED level 4 (non-tertiary post-secondary) or ISCED level 5 (vocational tertiary); and
- (4) ISCED level 6 (bachelor's level tertiary) or ISCED level 7 (master's level tertiary) or ISCED level 8 (doctorate level tertiary).

Immigrant background

Immigrant background (S_IMMIG) was created using the three responses to StQ 4, which collected information on the country of birth of the students and their parents. National research coordinators (NRCs) were asked to adapt the terms [Parent or guardian 1] and [Parent or guardian 2] to an appropriate term for parents in their national context. Instructions were provided to students on who to select as a parent for each corresponding option (including how to respond if they spend time with more than one set of parents and if they only spend time with one parent). For each student (S_SBORN), their first parent (S_P1BORN), and their second parent (S_P2BORN), a code of 1 was given if they were born in the country of the assessment while a score of 2 was assigned if they were born outside the country of assessment. The index S_IMMIG had the following three categories (in parentheses the code for each category in the dataset is presented):

- (0) Students and/or at least one parent born in country of test;
- (1) Student born in country of test but both/only parent(s) born abroad; and
- (2) Student and both/only parent(s) born abroad.

We assigned missing values to students with missing responses for either their own place of birth, or that of all parents, or for all three questions. The analyses of the relationship of immigrant background with computer and information literacy (CIL) and computational thinking (CT) were based on a dichotomous indicator variable that distinguished between students with immigrant background (coded 1) and students without immigrant background (coded 0) called S_IMMBGR.

Home language

Home language (S_TLANG) was obtained from StQ 5, where the students were asked if the language spoken at home most of the time was the language of assessment or another language;²⁶ this index had the following two categories:

- (0) The language spoken at home most of the time differed from the language of assessment; and

²⁶ Most countries collected more detailed information on language use. This information is not included in the international database.



- (1) The language spoken at home most of the time was the language of assessment.

Parents' or guardians' occupational status

Occupational status of the first parent or guardian (S_P1ISEI), occupational status of the second parent or guardian (S_P2ISEI), and the highest occupational status of both parents or guardians (S_HISEI), are three indices obtained from StQ 6 to StQ 12, where StQ 6 and 10 ask students whether their parents are in paid work or not while StQ 7/8 and 11/12 ask students to provide details about their parents' jobs, using pairs of questions with open text-response formats. The paid work status of the students' parents was classified into S_P1WORK (parent or guardian 1) and S_P2WORK (parent or guardian 2) (0 indicating that the parent is not in paid work, and 1 indicating that the parent is in paid work). The students' text responses were coded by national centers into four digit codes using the International Standard Classification of Occupations (ISCO-08) framework (International Labour Organization, 2007). These codes are contained in the indices S_P1ISCO and S_P2ISCO for the students' parents or guardians. We then mapped these codes to the international socioeconomic index of occupational status (S_P1ISEI and S_P2ISEI) (Ganzeboom et al., 1992), where higher ISEI scores indicate higher levels of occupational status.

Level of parental education

Highest educational level of parental education (S_HISCED) was created from StQ 9 and 13, which asked students about the highest parental education attainment of their parents, and provided the data for measuring another important family background variable. Levels of parental education were classified according to the ISCED levels.

Recoding educational qualifications into the following categories provided indices of highest parental educational attainment:

- (0) Did not complete ISCED level 2;
- (1) ISCED level 2 (lower-secondary);
- (2) ISCED level 3 (upper-secondary);
- (3) ISCED level 4 (non-tertiary post-secondary) or ISCED level 5 (vocational tertiary); and
- (4) ISCED level 6 (bachelor's level tertiary) or ISCED level 7 (master's level tertiary) or ISCED level 8 (doctorate level tertiary).

Indices with these categories are available for each student's first parent or guardian (S_P1ISCED) (Question 9) and second parent or guardian (S_P2ISCED) (Question 13). S_HISCED corresponds to the higher ISCED level of either parent.

Literacy resources at home

Students' home literacy resources (S_HOMLIT) was created from StQ 14, which asked students how many books they had in their homes. This index had the following categories:

- (0) 0 to 10 books;
- (1) 11 to 25 books;
- (2) 26 to 100 books;
- (3) 101 to 200 books; and
- (4) More than 200 books.

Internet connection at home

Internet connection at home (S_INTNET) was derived from the two items included in StQ 15 where students were asked about the quality of their internet connection at home. The first item was about



how often the internet in their home drops out and the second about how often the internet connection in their home is so slow that it made it difficult for them to complete their schoolwork. One response option for both items was “there is no internet connection in my home.” In case students gave this response for both items this was coded as 0. Any other combination of responses was coded as 1.

Computer experience in years

Computer experience in years (S_EXCOMP), includes desktop, laptop, notebook/netbook, and tablet devices, was created from StQ 17, which asked students about the number of years of experience in using computers. The index has the following categories:

- (0) Never or less than one year;
- (1) At least one year but less than three years;
- (2) At least three years but less than five years;
- (3) At least five years but less than seven years; and
- (4) Seven years or more.

ICT studies in current school year

ICT studies in current school year (S_ICTSTUD) was created from StQ 27 which asked students whether they studied a computing, computer science, information technology, informatics, or similar during the current school year where “Yes” was coded as 1 and “No” as 0.

Teacher questionnaire

Sex of teacher

Sex of teacher (T_SEX) was computed from the data captured from TQ 1. Female teachers were coded as 1, male teachers were coded as 0. As for students, countries could adapt the response categories to include one or more additional categories for teachers to self-identify as neither male nor female. Few teachers responded using the additional categories in countries where the national option was implemented, during data processing data in any additional categories were recoded to missing.

Teacher age

Teachers were asked to report their age in TQ2. The index T_AGE was created from TQ2. In case of missing information the data from the tracking form was used to calculate the age of the teacher.

Teaching in target grade schools in the current year

Teaching in target grade schools in the current year (T_WGT) was created from TQ 4 which asked teachers to indicate the number of schools in which they teach the target grade population for the current school year. The coding of the response categories are shown as below:

- (1.00) Only in the sampled school;
- (0.50) In the sampled school and another school;
- (0.33) In the sampled school and in two other schools; and
- (0.25) In the sampled school and in three or more other schools.

ICT experience with ICT use

ICT experience with ICT use during lessons (T_EXLES) and ICT experience with ICT use for preparing lessons (T_EXPREP) were created from Question 5 which asked teachers to indicate how long they have been using computers for preparing (TQ5a) or during lessons (TQ5b). These indices were coded as:

- (0) Never;
- (1) Less than two years;
- (2) At least two years but less than five years;
- (3) At least five years but less than ten years; and
- (4) Ten years or more.

School-level questionnaires

Sex of principal

Sex of principal (P_SEX) was created from PQ 22, which asked whether respondents were male or female where female principals were coded as 1, and male principals coded as 0. As for students and teachers, countries could adapt the response categories to include one or more additional categories for principals to self-identify as neither male nor female. Few principals responded using the additional categories in countries where the national option was implemented, during data processing data in any additional categories were recoded to missing.

Number of students in school and grade

Number of students in the entire school (P_NUMSTD) and the Number of students in the target grade (P_NUMTAR) were created from PQ 1 and PQ 2, which asked the principals about the number of female and male students in the entire school and also at the target grade. Categorical versions of the two indices were created. P_NUMSTD_CAT has the following values:

- (1) 1–300;
- (2) 301–600;
- (3) 601–900; and
- (4) More than 900.

Index P_NUMTAR_CAT has the following values:

- (1) 1–100;
- (2) 101–200; and
- (3) More than 200.

Number of grades in school

Number of grades in school (P_NGRADE) was created from PQ 3, which asked principals to report the lowest (youngest) grade and the highest (oldest) grade taught in their school. The difference between these two grades was calculated as the number of grades in school.

Number of teachers

Number of teachers (P_NUMTCH) was created from PQ 4, which asked the principals about the number of teachers in the school. This index was calculated by summing the total number of full time teachers with the total number of part-time teachers weighted at 50 percent. Further, the index was categorized and P_NUMTCH_CAT was created which has following categories:

- (1) 1–25;
- (2) 26–50;
- (3) 51–75; and
- (4) More than 75.



Ratio of school size and teachers

Ratio of school size and teachers (P_RATTCH) was calculated by dividing the number of teachers (P_NUMTCH) by the number of students (P_NUMSTD) in a school.

Private school indicator

Private school indicator (P_PRIV) was created from PQ 6a which collected information about whether the school was a public school or private school where public schools are coded as 0 and private schools coded as 1.

School composition by student background

School composition by student background (P_COMP) was created from PQ 6b which asked principals to estimate the percentage of students at their school coming from economically affluent homes and those from economically disadvantaged homes ("0–10%," "11–25%," "26–50%," and "more than 50%" for each home type). The index was coded as:

- (1) schools with more students from socio-economically affluent than disadvantaged homes;
- (2) schools with students from neither more socio-economically affluent nor more disadvantaged homes; and
- (3) schools with more students from socio-economically disadvantaged than affluent homes.

ICT experience in years in the school

ICT experience in years in the school (C_EXP) was created from CQ 3, which asked coordinators to indicate the number of years that the school has been using ICT for teaching and learning. This index was coded as:

- (0) Never, we do not use ICT;
- (1) Fewer than 5 years;
- (2) At least 5 but fewer than 10 years; and
- (3) 10 years or more.

Number of devices at the school

Question 7a of the ICT coordinator questionnaire collected data on the number of desktop computers, the number of laptops (including notebooks and netbooks), and the number of tablet devices in the school. The sum of ICT devices (C_ICTDEV) was derived by adding these numbers regardless if a response for all of them was provided. Respondents were also asked to indicate the number of these different types of devices that were available for student use. These were summed to derive an index of sum of ICT devices available for student use (C_ICTSTD), again not requiring a response to all of them. The Question 7b asked respondents to indicate the number of school-provided interactive whiteboards available in the school. In conjunction with the number of students at school (P_NUMSTD) these data also provided the following ratios:

- Ratio of school size and number of ICT devices (C_RATDEV) = Number of students in the school (P_NUMSTD) / number of ICT devices in the school altogether (C_ICTDEV).
- Ratio of school size and number of devices available for students (C_RATSTD) = Number of students in the school (P_NUMSTD) / number of ICT devices in the school that are available to students (C_ICTSTD).
- Ratio of school size and smart boards (C_RATSMB) = Number of students in the school (P_NUMSTD) / number of smart boards or interactive white boards available (II3G07B).

In ICILS 2018 the calculation of these indices was slightly different. This is why the indices are cal-

culated also with the procedure equivalent to 2018 and this is indicated in the index name. Indices C_ICTDEV_2018 and C_ICTSTD_2018 were calculated as the sum of all devices only, if all three responses were given. Following, the indices C_RATDEV_2018, C_RATSTD_2018 had to be calculated as well as described above.

12.2 Students' socioeconomic background indices

In ICILS 2023, two students' socioeconomic background indices were constructed: the national index of students' socioeconomic background (NISB) and the international index of students' socioeconomic background (IISB). Both indices were constructed using the same procedure. Nevertheless, the NISB was constructed individually for each country, only taking into account that country's data and the IISB was constructed using data from all countries that met participation requirements.

It is important to note that neither index should be used in absolute terms, their values do not indicate if a student has or has not a high or low socioeconomic background. Both indices only indicate the relative socioeconomic position of each student within their country, for the NISB, and within the total sample, for the IISB. Therefore, both indices can only be used within analyses that evaluate the relation between two or more variables, like correlation analysis and regression analysis. Moreover, when dealing with a single country analysis the NISB should be used, because their parameters are adjusted to the specific country, and when dealing with multiple countries, the IISB should be used because the NISB is not comparable between countries.

Both the NISB and the IISB were derived from the following three indices: highest occupational status of parents (S_HISEI), highest educational level of parents (S_HISCED), and the number of books at home (S_HOMLIT).

With these three indices, we constructed a three-by-three correlation matrix using the appropriate correlation coefficient for each index combination: Spearman's correlation coefficient between S_HISEI and S_HISCED, and between S_HISEI and S_HOMLIT, and a polychoric correlation coefficient between S_HISCED and S_HOMLIT. Later on, we conducted a principal component analysis (PCA) on this correlation matrix.

PCA was used to reduce the dimensionality of the original dataset while preserving as much variance as possible (Hastie et al., 2009; Jolliffe, 2002). The results of the first principal component, that explains the largest portion of the variance, were used as an indication of dimensionality. We expected that more than 50 percent of variance is explained by the first component in a unidimensional construct, like the students' socioeconomic background. As shown in Table 12.1 this is the case for all countries.



Table 12.1: Percentage of imputed data, reliability, explained variance, and component loadings of NISB by country and IISB

Country	Imputed data (%)	Cronbach's alpha	Explained variance	Component loadings		
				S_HISEI	S_HISCED	S_HOMLIT
Austria	3.7	0.70	0.62	0.79	0.81	0.76
Azerbaijan	13.5	0.54	0.52	0.74	0.75	0.66
Belgium (Flemish)	8.5	0.64	0.58	0.79	0.81	0.69
Bosnia and Herzegovina	3.3	0.67	0.61	0.80	0.85	0.68
[†] Chile	9.6	0.64	0.59	0.79	0.82	0.68
Chinese Taipei	2.0	0.67	0.61	0.76	0.82	0.75
Croatia	4.2	0.68	0.61	0.80	0.83	0.71
Cyprus	4.7	0.65	0.59	0.77	0.80	0.73
Czech Republic	2.2	0.68	0.61	0.79	0.83	0.73
Denmark	4.4	0.64	0.58	0.76	0.80	0.73
Finland	5.1	0.60	0.55	0.76	0.78	0.69
France	5.1	0.66	0.60	0.80	0.80	0.72
Germany	15.1	0.70	0.63	0.80	0.81	0.76
Greece	2.9	0.67	0.61	0.79	0.83	0.71
Hungary	3.0	0.79	0.71	0.83	0.88	0.80
Italy	2.8	0.69	0.62	0.80	0.83	0.71
Kazakhstan	3.5	0.50	0.50	0.75	0.76	0.61
Korea, Republic of	7.7	0.55	0.52	0.68	0.76	0.73
Kosovo	5.4	0.63	0.57	0.76	0.81	0.70
Latvia	15.8	0.65	0.59	0.80	0.81	0.69
Luxembourg	8.0	0.75	0.67	0.83	0.85	0.77
Malta	5.9	0.62	0.57	0.79	0.79	0.68
[†] Netherlands	12.3	0.61	0.56	0.76	0.78	0.71
Norway	5.6	0.64	0.58	0.75	0.78	0.74
Oman	8.4	0.48	0.50	0.75	0.80	0.54
Portugal	2.0	0.75	0.66	0.83	0.86	0.76
[†] Romania	10.6	0.73	0.65	0.81	0.85	0.75
Serbia	3.8	0.70	0.63	0.79	0.85	0.74
Slovak Republic	3.2	0.75	0.67	0.81	0.86	0.78
Slovenia	4.7	0.64	0.58	0.77	0.81	0.70
Spain	3.7	0.70	0.62	0.80	0.83	0.74
Sweden	8.1	0.66	0.59	0.74	0.80	0.77
[†] United States	16.3	0.70	0.62	0.79	0.83	0.75
Uruguay	4.7	0.68	0.61	0.77	0.82	0.74
[†] North Rhine- Westphalia, Germany	19.8	0.70	0.62	0.79	0.80	0.78
IISB	6.3	0.67	0.60	0.79	0.82	0.73

[†] These countries were excluded from the international pooled dataset.



Before computing their score we imputed values for students who had missing data for one of the three indicators, we used predicted values plus a random component based on a regression on the other two variables that had been estimated for students with values on all three variables. This imputation procedure was carried out for each national sample separately. The percentage of imputed data by country is shown in Table 12.1.

With the Eigenvectors scaled up by the variances, i.e., component loadings, from the original data and the imputed data we can calculate component scores for each student using the following formula:

$$\sum_{i=1}^3 \frac{L_i}{\lambda_1} \left(\frac{x_{ij} - \bar{x}_i}{\sigma_{x_i}} \right),$$

where x_{ij} is the value of student j for the index i (S_HISEI, S_HISCED, or S_HOMLIT), L_i is the component loading of index i , and λ_1 is the first eigenvalue. As we have an overall IISB, and a NISB for each country, this score was calculated for each country twice: one using the pooled statistics and parameters, and one using the own country statistics and parameters. Finally, we standardized each index with a mean of 0 (pooled mean for IISB and country mean for NISB) and a standard deviation of 1.

12.3 Scaling procedures

Many of the ICILS questionnaire questions were created with sets of items that were developed specifically with the purpose and assumption that the data from the items within each set could be used to establish measurement scales of a latent trait underpinning the item sets. These were developed with reference to the ICILS 2023 assessment framework (see Fraillon & Rožman, 2024). The questionnaire item development process is described in Chapter 4.

In this section, we explain the processes used to evaluate the viability of establishing the planned scales from sets of items. These processes evaluate both the degree to which individual items within a set contributed to the scale measures and the measurement properties of the derived scales. The majority of proposed scales were to be derived from sets of items with four Likert-type response options.

Review of item statistics

Item reviews began with an evaluation of item-level non-response. In the ICILS 2023 questionnaires, there are three types of item level-non response.

- *Omitted*: The respondent had a chance to answer the item but did not do so, leaving the corresponding item without a response.
- *Not administered*: This signified that the item or question was not administered to the respondent, meaning that the respondent could not read and answer the question. This could happen for example in the case where a country chose not to administer an item or question that was an international option.
- *Logically not applicable*: This missing code was assigned to questionnaire items that depended on a filter question. If a respondent's answer to the filter question indicated that subsequent questions were not relevant, those follow-up items were coded as 'logically not applicable.'

Omitted codes were considered only when presenting the percentage distributions and did not contribute to the calculation of scale statistics. *Not administered* and *logically not applicable* codes did not contribute to the calculation of the item statistics or scale construction.

We then examined Cronbach's alpha measure of the *internal consistency* or *reliability* of each scale. When reviewing Cronbach's alpha, we regarded values above 0.7 as indicating satisfactory reliability and those above 0.8 as showing high degrees of reliability (see, for example, Nunnally & Bernstein, 1994, pp. 264–265). The reliabilities of all ICILS 2023 questionnaire scale indices are presented in H.



Scale check

After reviewing item statistics to select the proposed item content for each scale, we completed further checks to determine whether the item data were suitable for generating IRT scales.

Only proposed scales comprising data from three or more items were considered for further scale construction (using IRT scaling) and scoring.

In addition, the following two criteria were applied to the data for item sets in the proposed scales to determine their fitness for further scaling analyses. These criteria were first applied to the item sets using the pooled international data.

- There must be at least 200 valid cases (for a case to be considered valid, the respondent has to have provided responses to at least two items in the proposed scale).
- There must be variation shown in the responses.

For proposed item sets that met the above criteria using the pooled international data, the same criteria were then applied separately for data within each country. Where the data within a country did not meet one or more criteria for a proposed scale, the data from this country were not included into multi-group IRT analysis for that scale.

It is noted that even though some countries were not included into by country analysis due to failure to meet all criteria, all countries received scale scores for valid cases, using the international parameters.

We checked as well, if Cronbach's alpha measure of the *internal consistency* or *reliability* of each scale was low based on criteria described above. All ICILS 2023 questionnaire scaled indices presented satisfactory reliability, further supporting the internal consistency of each scale.

Scale dimensionality

Since IRT models often assume unidimensionality (Embretson & Reise, 2000; Lord, 1980), PCA was conducted first to evaluate the dimensionality of the data for each proposed scale, and how much variance the first principal component accounted for. The steps to determine the dimensionality of the original dataset are shown as below (Jolliffe, 2002):

1. Calculate the polychoric correlation matrix with all scale items;
2. Calculate the Eigenvectors and Eigenvalues from the polychoric correlation matrix;
3. Sort the Eigenvalues associated with each principal component in descending order;
4. Determine the number of principal components.

A scree plot is used to choose the number of principal components (Cattell, 1966). It is a visual representation of how much variance each principal component explains with X axis indicates the principal components in descending order and Y axis shows the Eigenvalues or the percent of variance explained. The point where the curve starts to flatten (the so called "elbow point" or "inflection point") is used to select the number of principal components. Components beyond this point explain significantly less variance within the data.

Item response modeling

For ICILS 2023, the majority of the context questionnaire items are polytomous, where each item has multiple ordered response categories (e.g., rating scales), then Partial Credit Modeling (PCM), an extension of Rasch model was applied (Masters & Wright, 1997). The mathematical formulations of both Rasch model and PCM model are shown as below:

$$P_i(\theta_n) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)},$$



where $P_i(\theta_n)$ is the probability of person n scoring 1 on item i , θ_n is the estimated latent trait of person n , and δ_i is the estimated location of item i on this dimension. For each item, item responses are modeled as a function of the latent trait θ_n :

$$P_{X_i}(\theta_n) = \frac{\exp\left(\sum_{j=0}^x (\theta_n - \delta_i + \tau_{ij})\right)}{\sum_{h=0}^{m_i} \exp\left(\sum_{j=0}^h (\theta_n - \delta_i + \tau_{ij})\right)} \quad x_i = 0, 1, \dots, m_i,$$

where $P_{X_i}(\theta_n)$ denotes the probability of person n scoring x on item i , θ_n denotes the person's latent trait, the item parameter δ_i describes the location of the item on the latent continuum, and τ_{ij} provides an additional step parameter.

International parameter calibration

The international item parameters were derived from two PCM models shown below using equally weighted national datasets:

A) International Pooled PCM Model.

The first PCM model was constructed with international pooled dataset with senate weights so that each country contributes equally to the calibration of the item parameters. Data were included only from countries meeting the sampling participation requirements relevant to the specific questionnaire.²⁷

B) Multiple Group PCM Model.

Then the item parameters obtained from the pooled model were used as initial parameters in the multiple group PCM model. It is noted that countries that did not meet the previously described criteria (sufficient number of valid cases or variation for all items) were not included in this model.

Further, the item parameters obtained from the multiple group PCM were fixed to build PCM models for each country, including those countries that were excluded from the multiple group PCM so that all country participants could have scale scores except for invalid cases.

Due to the smaller sample size at the school level, only two groups (the Czech Republic and Spain) fulfilled the criteria for inclusion in the multiple-group model. For some scales derived from the principal and coordinator questionnaires, there were fewer than two groups available for multiple-group analyses. In these cases, IRT parameters from the international pooled PCM model were fixed and used to derive scale scores for each country.

Model evaluation and scoring

Once the PCM models were constructed, the following statistics were used to evaluate the performance of the models:

- Inlier-sensitive or Information-weighted Fit (*infit*): This is an item statistic to measure the discrepancy between the observed data for the item and the IRT model implied item value. As a general rule of thumb, 1 indicates a perfect fit; values less than 1 suggest overfit, values bigger than 1 suggest underfit, meaning that there is more variation in the item than the model predicts, excluding outliers (Wright & Masters, 1982).
- Outlier-sensitive Fit (*outfit*): This is a similar item statistic as *infit* to measure the discrepancy between the observed data for the item and the IRT model implied value. It also uses 1 to indicate

²⁷ Romania met the sampling participation requirements, but due to late testing their data were not included in the final calibration.



the perfect fit; but unlike infit, outfit considers all responses, including the outliers (Wright & Masters, 1982).

- Weighted Root Mean Square Deviation (WRMSD): This metric assesses the weighted average of the squared differences between observed and expected responses. The weighting usually takes into account the standard errors of the observed scores, making it more sensitive to areas with higher measurement precision. A lower WRMSD value indicates a better fit between the model predictions and the observed data.
 - WRMSD < 0.1: Excellent fit;
 - 0.1 ≤ WRMSD < 0.2: Good fit;
 - 0.2 ≤ WRMSD < 0.3: Moderate fit;
 - WRMSD ≥ 0.3: Poor fit, suggesting the model may not be adequately describing the data.
- Root Mean Square Deviation (RMSD): This is the square root of the average of the squared differences between observed and expected responses. It provides a measure of the overall fit of the model. Similar to WRMSD, a lower RMSD value indicates a better fit. While specific cutoffs are less common, an RMSD cutoff value of 0.3 has shown to be effective in simulation studies (Buchholz & Hartig, 2019) and was applied in other ILSAs (Yamamoto et al., 2013). In this report, we use 0.25 as the cutoff point.

Following the estimation of international item parameter from the calibration sample, we computed weighted likelihood estimation to obtain individual student, teacher, or school scores. Weighted likelihood estimations are computed by minimizing the equation:

$$\sum_{i \in \Omega} \left[\left(r_x + \frac{J_n}{2I_n} \right) - \sum_{i=1}^k \frac{\exp \left(\sum_{j=0}^x (\theta_n - \delta_i + \tau_{ij}) \right)}{\sum_{h=0}^{m_i} \exp \left(\sum_{j=0}^h (\theta_n - \delta_i + \tau_{ij}) \right)} \right] = 0,$$

for each case n , where r_x is the sum score obtained from a set of k items with j steps between adjacent categories. This can be achieved by applying the Newton-Raphson method. The term $J_n/2I_n$ (with I_n being the information function for student n and J_n being its derivative with respect to q) is used as a weight function to account for the bias inherent in maximum likelihood estimation (Warm, 1989). We used the R library TAM (Robitzsch et al., 2024) for the pre-calibration of item parameters in order to subsequently derive scale scores.

For all questionnaire scales in ICILS 2023 the transformation of weighted likelihood estimates to an international metric resulted in reporting scales with an ICILS 2023 average of 50 and a standard deviation of 10 for equally weighted datasets from the countries that met sample participation requirements. This is achieved by applying the following formula:

$$\theta'_n = 50 + 10 \frac{\theta_n - \mu_\theta}{\sigma_\theta}$$

where θ'_n are the scores in the international metric, θ_n are the original weighted likelihood estimates in logits, and μ_θ is the international mean of logit scores with equally weighted country subsamples, and σ_θ is the corresponding international standard deviation of the original weighted likelihood estimates.

12.4 Scaling results

In this section we present the results for ICILS 2023 questionnaire scales. For each scale, we present a short description, its items, their response categories, and their item parameters. This section is divided into: student scales, teacher scales, principal scales, and ICT coordinator scales.

Student Scales

Learning about internet related tasks at school (S_LRNINTS)

In Question 20a.i, students were asked to indicate the extent to which they had learned various internet-related tasks at school. For each task, they selected from the following response options: “To a large extent,” “To a moderate extent,” “To a small extent,” or “Not at all.” The responses to these items were used to derive a scale reflecting students’ learning about internet-related tasks at school (S_LRNINTS). This scale includes items related to students’ ability to search for information online, refine search results, evaluate the reliability of information, include accurate references, and recognize potential online scams or deceptive messages.

Higher values on this scale indicate greater self-reported learning of internet-related skills through formal school instruction. The scale is composed of 6 items, item wording is presented in Table 12.2. The IRT results are good (Table 12.3). Both outfit and infit are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.2: Item wording for the scale Learning about internet related tasks at school (S_LRNINTS)

Scale or item	Question or item wording
S_LRNINTS	To what extent have you learned how to do the following Internet-related tasks at school and outside of school? (At school) <i>Recoded response categories for scaling: 0=Not at all;1=To a small extent;2=To a moderate extent;3=To a large extent</i>
IS3G20AA1	Use the internet to find information (e.g., by using e websites, databases, archives, digital libraries, search engines)
IS3G20AB1	Refine internet searches, so the results better match what you are looking for
IS3G20AC1	Evaluate the reliability (trustworthiness) of information on the internet
IS3G20AD1	Include accurate references to internet sources
IS3G20AE1	Judge whether a message from someone is a scam (e.g., a message that tricks you into downloading a virus)
IS3G20AF1	Manage privacy settings for internet accounts and ICT devices (e.g., control who can contact you and what information about you is shared with advertising companies)

Table 12.3: Item parameters, item fits, and WRMSD for the scale Learning about internet related tasks at school (S_LRNINTS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G20AA1	-0.185	-1.414	-0.004	1.418	0.978	0.984	0.071
IS3G20AB1	0.017	-1.128	-0.210	1.338	0.947	0.962	0.047
IS3G20AC1	-0.096	-1.009	-0.209	1.217	0.972	0.981	0.053
IS3G20AD1	-0.162	-0.997	-0.242	1.240	1.031	1.041	0.061
IS3G20AE1	0.136	-0.427	-0.146	0.572	1.064	1.071	0.060
IS3G20AF1	0.289	-0.664	-0.178	0.842	1.056	1.065	0.061

Learning about internet related tasks outside of school (S_LRNINTO)

In Question 20a.ii, students were asked to indicate the extent to which they had learned various internet-related tasks outside of school. For each task, they selected from the following response options: “To a large extent,” “To a moderate extent,” “To a small extent,” or “Not at all.” The responses to these items were used to derive a scale reflecting students’ learning about internet-related tasks outside of school (S_LRNINTO). This scale includes items related to students’ ability to search for information



online, refine search results, evaluate the reliability of information, include accurate references, and recognize potential online scams or deceptive messages.

Higher values on this scale indicate greater self-reported learning of internet-related skills through informal means outside of school settings. The scale is composed of 6 items, item wording is presented in Table 12.4. The IRT results are good (Table 12.5). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.4: Item wording for the scale *Learning about internet related tasks outside of school* (S_LRNINTO)

Scale or item	Question or item wording
S_LRNINTO	To what extent have you learned how to do the following Internet-related tasks at school and outside of school? (Outside of school) <i>Recoded response categories for scaling: 0=Not at all;1=To a small extent;2=To a moderate extent;3=To a large extent</i>
IS3G20AA2	Use the internet to find information (e.g., by using e websites, databases, archives, digital libraries, search engines)
IS3G20AB2	Refine internet searches, so the results better match what you are looking for
IS3G20AC2	Evaluate the reliability (trustworthiness) of information on the internet
IS3G20AD2	Include accurate references to internet sources
IS3G20AE2	Judge whether a message from someone is a scam (e.g., a message that tricks you into downloading a virus)
IS3G20AF2	Manage privacy settings for internet accounts and ICT devices (e.g., control who can contact you and what information about you is shared with advertising companies)

Table 12.5: Item parameters, item fits, and WRMSD for the scale *Learning about internet related tasks outside of school* (S_LRNINTO)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G20AA2	-0.449	-0.908	-0.122	1.030	0.983	1.006	0.050
IS3G20AB2	-0.047	-0.948	-0.195	1.143	0.972	0.993	0.045
IS3G20AC2	0.167	-1.017	-0.179	1.196	0.972	0.986	0.048
IS3G20AD2	0.403	-0.914	-0.224	1.138	1.037	1.044	0.060
IS3G20AE2	-0.058	-0.289	-0.172	0.461	1.080	1.086	0.067
IS3G20AF2	-0.016	-0.515	-0.230	0.745	1.028	1.040	0.053

Learning about safe and responsible ICT use at school (S_LRNSAFE)

In Question 20c, students were asked to indicate the extent to which they had learned about various aspects of safe and responsible ICT use at school. For each topic, they selected from the following response options: “To a large extent,” “To a moderate extent,” “To a small extent,” or “Not at all.” The responses to these items were used to derive a scale reflecting students’ learning about safe and responsible ICT use at school (S_LRNSAFE). This scale includes items related to students’ knowledge of responsible and respectful use of social media, recognizing cyberbullying, understanding the physical health implications of ICT use, and awareness of psychological well-being related to ICT use.

Higher values on this scale indicate greater self-reported learning about these topics through school instruction. The scale is composed of 4 items, item wording is presented in Table 12.6. The IRT results are very good (Table 12.7). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.6: Item wording for the scale Learning about safe and responsible ICT use at school (S_LRNSAFE)

Scale or item	Question or item wording
S_LRNSAFE	To what extent have you learned about the following topics at school? <i>Recoded response categories for scaling: 0=Not at all;1=To a small extent;2=To a moderate extent;3=To a large extent</i>
IS3G20CA	Responsible and respectful use of social media (including the use of images and personal information)
IS3G20CB	How to recognize cyberbullying
IS3G20CD	Physical health and ICT use
IS3G20CE	Psychological health and ICT use

Table 12.7: Item parameters, item fits, and WRMSD for the scale Learning about safe and responsible ICT use at school (S_LRNSAFE)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G20CA	-0.342	-1.258	-0.296	1.554	1.052	1.061	0.061
IS3G20CB	-0.155	-1.065	-0.239	1.304	1.075	1.083	0.064
IS3G20CD	0.153	-1.333	-0.211	1.544	0.867	0.900	0.060
IS3G20CE	0.343	-1.211	-0.158	1.369	0.870	0.905	0.059

Academic-media multitasking (S_ACMULT)

Question 21 of the student questionnaire asked students to indicate the frequency with which they use ICT for various activities outside of school. The response options included: “Never,” “Almost never,” “Sometimes,” “Often,” and “Very often.” A subset of items from this question was used to derive a scale reflecting students’ engagement in academic-media multitasking. The S_ACMULT scale measures the extent to which students engage in concurrent academic and media-related activities using ICT. It includes items such as text chatting, checking social media for updates, using the internet to find information, and watching online videos.

Higher values on this scale indicate more frequent engagement in multitasking behaviors while using ICT. The scale is composed of 6 items, item wording is presented in Table 12.8. The IRT results are good (Table 12.9). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.8: Item wording for the scale Academic-media multitasking (S_ACMULT)

Scale or item	Question or item wording
S_ACMULT	Outside of school, how often do you do the following activities not related to your schoolwork at the same time as doing your schoolwork? <i>Recoded response categories for scaling: 0=Never;1=Almost never;2=Sometimes;3=Often;4=Very often</i>
IS3G21A	Text chat with others (with any device including smartphones)
IS3G21B	Use social media (e.g., [Instagram, TikTok and Snapchat]) to post or view content
IS3G21C	Check social media for new posts or responses to my posts
IS3G21D	Use the Internet to find information about things that interest me
IS3G21E	Watch online videos, live streams or television (e.g., [YouTube, Twitch, Netflix])
IS3G21F	Listen to music, podcasts, or the radio (on any device)



Table 12.9: Item parameters, item fits, and WRMSD for the scale Academic-media multitasking (S_ACMULT)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	Outfit	Infit	WRMSD
IS3G21A	-0.118	-0.399	-0.903	0.437	0.866	0.889	0.907	0.056
IS3G21B	0.058	-0.243	-0.518	0.282	0.479	0.876	0.927	0.059
IS3G21C	0.523	-0.633	-0.629	0.509	0.752	0.987	1.006	0.054
IS3G21D	-0.110	-0.820	-0.850	0.437	1.233	1.121	1.110	0.050
IS3G21E	-0.065	-0.429	-0.575	0.320	0.684	1.056	1.055	0.043
IS3G21F	-0.287	-0.324	-0.506	0.414	0.415	1.229	1.179	0.041

General classroom applications (S_GENCLASS)

Two scales were constructed from the statements in Question 23. Students were asked to indicate how often they used different types of ICT applications during their lessons throughout the school year. For each application, they selected from the following response options: “Never,” “In some lessons,” “In most lessons,” or “In every lesson.” The responses to these items were used to derive a scale reflecting students’ use of general ICT applications in the classroom (S_GENCLASS). This scale includes items related to students’ use of word-processing software, presentation software, spreadsheets, and computer-based information resources (e.g., websites and online encyclopedias) in their schoolwork.

Higher values on this scale indicate more frequent use of general ICT applications during classroom activities. The scale is composed of 4 items, item wording is presented in Table 12.10. The IRT results are very good (Table 12.11). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.10: Item wording for the scale General classroom applications (S_GENCLASS)

Scale or item	Question or item wording
S_GENCLASS	When studying throughout this school year, how often did you use the following tools during lessons? <i>Recoded response categories for scaling: 0=Never;1=In some lessons;2=In most lessons;3=In every or almost every lesson</i>
IS3G23A	Word-processing software (e.g., [Microsoft Word, Apple Pages, Google Docs])
IS3G23B	Presentation software (e.g., [Microsoft PowerPoint, Apple Keynote, Google Slides])
IS3G23C	Spreadsheets (e.g., [Microsoft Excel, Apple Numbers, Google Sheets])
IS3G23H	Computer-based information resources (e.g., websites, wikis, encyclopaedia)

Table 12.11: Item parameters, item fits, and WRMSD for the scale General classroom applications (S_GENCLASS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G23A	-0.184	-2.010	0.546	1.464	0.901	0.920	0.083
IS3G23B	-0.298	-2.458	0.482	1.976	0.880	0.884	0.068
IS3G23C	0.535	-1.853	0.412	1.441	1.047	1.079	0.085
IS3G23H	-0.053	-1.930	0.325	1.605	1.093	1.096	0.064

Specialist classroom applications (S_SPECLASS)

Another scale was constructed from the statements in Question 23. Students were asked to indicate



how often they used different types of ICT applications during their lessons throughout the school year. For each application, they selected from the following response options: “Never,” “In some lessons,” “In most lessons,” or “In every lesson.” The responses to these items were used to derive a scale reflecting students’ use of specialist ICT applications in the classroom (S_SPECLASS). This scale includes items related to students’ use of multimedia production tools (e.g., video editing software), concept mapping software, real-world data capture tools (e.g., speech recognition, data loggers), simulations and modeling software, and interactive digital learning resources.

Higher values on this scale indicate more frequent use of specialist ICT applications during classroom activities. The scale is composed of 8 items, item wording is presented in Table 12.12. The IRT results are good (Table 12.13). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for almost all items and all countries are below .25 except for Denmark who has relatively higher RMSD (0.250) for IS3G23E, indicating that this scale is invariant across countries except for Denmark.

Table 12.12: Item wording for the scale Specialist classroom applications (S_SPECLASS)

Scale or item	Question or item wording
S_SPECLASS	When studying throughout this school year, how often did you use the following tools during lessons? <i>Recoded response categories for scaling: 0=Never;1=In some lessons;2=In most lessons;3=In every or almost every lesson</i>
IS3G23D	Multimedia production tools (e.g., video editing, audio/music mixing, animation)
IS3G23E	Concept mapping software (e.g., [Inspiration, Webspiration])
IS3G23F	Tools that capture real-world data (e.g., speed, temperature) digitally for analysis
IS3G23G	Simulations and modelling software (e.g., physics simulators)
IS3G23I	Interactive digital learning resources (e.g., learning games or apps)
IS3G23J	Drawing and graphic design software (e.g., logo design, poster design, character illustration)
IS3G23L	A video conferencing system (e.g., [Zoom, WebEx GoTo Meeting, Google Meet])
IS3G23M	A computer programming/coding environment (e.g., [Python, LUA, Javascript, Scratch])

Table 12.13: Item parameters, item fits, and WRMSD for the scale Specialist classroom applications (S_SPECLASS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G23D	-0.074	-1.680	0.565	1.115	0.960	0.952	0.061
IS3G23E	0.404	-1.234	0.238	0.996	0.822	0.857	0.081
IS3G23F	0.235	-1.532	0.390	1.143	0.868	0.870	0.066
IS3G23G	0.272	-1.335	0.274	1.061	0.852	0.869	0.069
IS3G23I	-0.473	-2.124	0.532	1.592	1.093	1.080	0.072
IS3G23J	-0.065	-1.675	0.489	1.186	0.955	0.959	0.062
IS3G23L	-0.204	-1.463	0.369	1.094	1.181	1.138	0.079
IS3G23M	-0.093	-1.333	0.394	0.939	1.091	1.078	0.091

General ICT self-efficacy (S_GENEFF)

Two scales were constructed from the statements in Question 24. Students were asked to indicate how well they could perform various ICT-related tasks. For each task, they selected from the following response options: “Very well,” “Moderately well,” “I have never done this, but I could work out how to do it,” or “I do not think I could do this.” The responses to these items were used to derive a



scale reflecting students' general ICT self-efficacy (S_GENEFF). This scale includes items related to students' confidence in performing ICT tasks such as editing digital photographs, writing or editing text for a school assignment, searching for relevant information, changing device settings, and creating multimedia presentations.

Higher values on this scale indicate higher self-efficacy in using ICT for various tasks. The scale is composed of 10 items, item wording is presented in Table 12.14. The IRT results are acceptable (Table 12.15). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.14: Item wording for the scale General ICT self-efficacy (S_GENEFF)

Scale or item	Question or item wording
S_GENEFF	How well can you do each of these tasks when using ICT? <i>Recoded response categories for scaling: 0=I do not think I could do this;1=I have never done this, but I could work out how to do it;2=Moderately well;3=Very well</i>
IS3G24A	Edit digital photographs or other graphic images
IS3G24B	Write or edit text for a school assignment
IS3G24C	Search for relevant information for a school project on the internet
IS3G24E	Change the settings on a device to suit your needs and preferences
IS3G24F	Create a multi-media presentation (with sound, pictures, or video)
IS3G24G	Upload text, images, or video to an online profile
IS3G24H	Insert an image into a document or message
IS3G24I	Install a program or app
IS3G24J	Judge whether you can trust information you find on the internet
IS3G24M	Find the original sources of information referred to in an article on the internet, if the URL is not given

Table 12.15: Item parameters, item fits, and WRMSD for the scale General ICT self-efficacy (S_GENEFF)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G24A	0.159	-1.147	-0.724	1.871	1.100	1.101	0.062
IS3G24B	-0.378	-0.812	-0.938	1.749	0.858	0.877	0.068
IS3G24C	-0.495	-0.766	-0.828	1.594	0.857	0.898	0.065
IS3G24E	0.272	-1.100	-0.393	1.493	0.991	1.007	0.060
IS3G24F	0.379	-1.353	-0.260	1.613	1.004	1.016	0.068
IS3G24G	-0.164	-1.034	-0.268	1.302	0.899	0.929	0.066
IS3G24H	-0.492	-0.704	-0.535	1.239	0.770	0.843	0.081
IS3G24I	-0.397	-0.667	-0.287	0.954	1.008	1.004	0.069
IS3G24J	0.161	-1.038	-0.885	1.923	1.020	1.028	0.053
IS3G24M	0.955	-1.398	-0.255	1.653	1.246	1.242	0.076

Specialist ICT self-efficacy (S_SPECEFF)

This scale included items from Question 24 as well. Students were asked to indicate how well they could perform various specialist ICT-related tasks. For each task, they selected from the following response options: "Very well," "Moderately well," "I have never done this, but I could work out how to do it," or "I do not think I could do this." The responses to these items were used to derive a scale reflecting students' specialist ICT self-efficacy (S_SPECEFF). This scale includes items related to

students' confidence in performing specialized ICT tasks, such as building or editing a webpage, using a text-based programming language (e.g., Python, Java), and using visual coding platforms (e.g., Alice, GameMaker, Kodu, Scratch).

Higher values on this scale indicate greater self-efficacy in performing specialist ICT-related tasks. The scale is composed of 3 items, item wording is presented in Table 12.16. The IRT results are good (Table 12.17). Both outfit and infit are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.16: Item wording for the scale Specialist ICT self-efficacy (S_SPECEFF)

Scale or item	Question or item wording
S_SPECEFF	How well can you do each of these tasks when using ICT? <i>Recoded response categories for scaling: 0=I do not think I could do this;1=I have never done this, but I could work out how to do it;2=Moderately well;3=Very well</i>
IS3G24D	Build or edit a webpage
IS3G24K	Use a text-based programming language (e.g., [Python, JavaScript, Lua, Swift]) to write a simple computer program
IS3G24L	Use visual coding (e.g., [Alice, GameMaker, Kodu, Lego Mindstorms, MIT App Inventor, Scratch]) to develop a program that someone else could use

Table 12.17: Item parameters, item fits, and WRMSD for the scale Specialist ICT self-efficacy (S_SPECEFF)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G24D	-0.150	-1.732	0.064	1.668	1.058	1.063	0.054
IS3G24K	0.089	-1.638	-0.020	1.658	0.875	0.887	0.062
IS3G24L	0.062	-1.553	-0.121	1.674	0.930	0.939	0.065

Attitudes toward ICT - positive outcomes general (S_ICTPOSG)

Two scales were constructed from the statements in Question 25a. Students were asked to indicate their level of agreement with various statements about the benefits of ICT. For each statement, they selected from the following response options: "Strongly agree," "Agree," "Disagree," or "Strongly disagree." The responses to these items were used to derive a scale reflecting students' positive perceptions of ICT (S_ICTPOSG). This scale includes items related to beliefs that advances in ICT improve living conditions, help people understand the world better, bring social benefits, and provide value to society.

Higher values on this scale indicate stronger positive perceptions of ICT and its societal benefits. The scale is composed of 4 items, item wording is presented in Table 12.18. The IRT results are very good (Table 12.19). Both outfit and infit are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.18: Item wording for the scale Attitudes toward ICT - positive outcomes general (S_ICTPOSG)

Scale or item	Question or item wording
S_ICTPOSG	How much do you agree or disagree with the following statements about ICT? <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IS3G25AA	Advances in ICT usually improve people's living conditions
IS3G25AB	ICT helps us to understand the world better
IS3G25AF	ICT is valuable to society
IS3G25AG	Advances in ICT bring many social benefits

Table 12.19: Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - positive outcomes general (S_ICTPOSG)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G25AA	-0.036	-1.953	-0.887	2.840	0.948	0.964	0.056
IS3G25AB	-0.102	-2.018	-0.713	2.731	0.914	0.932	0.053
IS3G25AF	0.092	-1.939	-0.668	2.607	0.941	0.956	0.062
IS3G25AG	0.045	-1.966	-0.615	2.581	0.948	0.966	0.054

Attitudes toward ICT - negative outcomes (S_ICTNEG)

Another scale was constructed from the items in Question 25a. Students were asked to indicate their level of agreement with various statements about possible negative consequences of ICT. For each statement, they selected from the following response options: "Strongly agree," "Agree," "Disagree," or "Strongly disagree." The responses to these items were used to derive a scale reflecting students' negative perceptions of ICT (S_ICTNEG). This scale includes items related to concerns that ICT may increase social isolation, reduce job opportunities, lead to excessive screen time, and pose health risks.

Higher values on this scale indicate stronger negative perceptions of ICT and its societal impacts. The scale is composed of 4 items, item wording is presented in Table 12.20. The IRT results are very good (Table 12.21). Both outfit and infit are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.20: Item wording for the scale Attitudes toward ICT - negative outcomes (S_ICTNEG)

Scale or item	Question or item wording
S_ICTNEG	How much do you agree or disagree with the following statements about ICT? <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IS3G25AC	Using ICT makes people more isolated in society
IS3G25AD	With more ICT there will be fewer jobs
IS3G25AE	People spend far too much time using ICT
IS3G25AH	Using ICT may be dangerous for people's health



Table 12.21: Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - negative outcomes (S_ICTNEG)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G25AC	0.075	-1.947	-0.096	2.043	0.995	1.000	0.054
IS3G25AD	0.452	-1.876	0.062	1.814	1.072	1.069	0.052
IS3G25AE	-0.477	-1.372	-0.417	1.789	0.932	0.956	0.057
IS3G25AH	-0.051	-1.307	-0.462	1.769	1.001	1.005	0.047

Attitudes toward ICT - expectations for future use of ICT (S_ICTFUT)

Two scales were constructed from the statements in Question 25b. Students were asked to indicate their level of agreement with various statements regarding the role of ICT in their future studies and careers. For each statement, they selected from the following response options: “Strongly agree,” “Agree,” “Disagree,” or “Strongly disagree.” The responses to these items were used to derive a scale reflecting students’ expectations for ICT use in the future (S_ICTFUT). This scale includes items related to students’ aspirations to study ICT-related subjects, use ICT as an important part of their future work, engage in programming, and recognize the value of ICT skills for career prospects.

Higher values on this scale indicate stronger expectations for using ICT in future academic or professional settings. The scale is composed of 5 items, item wording is presented in Table 12.22. The IRT results are good (Table 12.23). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.22: Item wording for the scale Attitudes toward ICT - expectations for future use of ICT (S_ICTFUT)

Scale or item	Question or item wording
S_ICTFUT	How much do you agree or disagree with the following statements about ICT and school? <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IS3G25BC	I would like to study subjects related to ICT after [secondary school]
IS3G25BD	I hope that using ICT is a very important part of my future job
IS3G25BE	I hope that my future job involves programming
IS3G25BF	Learning how to use ICT applications will help me to do the work I am interested in
IS3G25BH	Learning how to use ICT well will help me get a well-paid job

Table 12.23: Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - expectations for future use of ICT (S_ICTFUT)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G25BC	0.328	-1.886	0.002	1.885	0.901	0.916	0.065
IS3G25BD	-0.052	-2.021	-0.149	2.169	0.860	0.879	0.064
IS3G25BE	0.719	-1.657	-0.006	1.664	0.918	0.938	0.065
IS3G25BF	-0.210	-1.862	-0.426	2.288	0.898	0.921	0.063
IS3G25BH	-0.785	-2.003	-0.487	2.490	1.168	1.201	0.070

Attitudes toward ICT - positive outcomes (S_ICTPOSS)

Another scale was constructed from the items in Question 25b. Students were asked to indicate their level of agreement with various statements regarding the role of ICT in learning. For each statement,

they selected from the following response options: “Strongly agree,” “Agree,” “Disagree,” or “Strongly disagree.” The responses to these items were used to derive a scale reflecting students’ perceptions of the benefits of ICT for learning (S_ICTPOSS). This scale includes items related to ICT making learning more enjoyable, the importance of learning how to use ICT, the significance of learning programming, and the need for students to stay updated with ICT advancements.

Higher values on this scale indicate stronger positive attitudes toward the role of ICT in education and its relevance for students’ learning and future opportunities. The scale is composed of 4 items, item wording is presented in Table 12.24. The IRT results are very good (Table 12.25). Both outfit and infit are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.24: Item wording for the scale Attitudes toward ICT - positive outcomes (S_ICTPOSS)

Scale or item	Question or item wording
S_ICTPOSS	How much do you agree or disagree with the following statements about ICT and school? <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IS3G25BB	Using ICT at school makes learning more fun
IS3G25BG	It is important for students to learn how to use ICT at school
IS3G25BI	It is important for students to learn programming at school
IS3G25BJ	It is important for students to keep up to date with changes in ICT

Table 12.25: Item parameters, item fits, and WRMSD for the scale Attitudes toward ICT - positive outcomes (S_ICTPOSS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G25BB	-0.160	-1.603	-0.599	2.203	1.053	1.062	0.049
IS3G25BG	-0.272	-1.315	-0.895	2.210	0.894	0.923	0.063
IS3G25BI	0.360	-1.858	-0.360	2.218	1.019	1.022	0.073
IS3G25BJ	0.072	-1.523	-0.743	2.266	0.931	0.946	0.052

Learning about CT in class (S_CTCLS)

Question 26 of the student questionnaire asked students to indicate the extent to which they had learned specific computational thinking (CT) skills in class. The response options for each item were: “To a large extent,” “To a moderate extent,” “To a small extent,” or “Not at all.” A set of items from this question was used to derive a scale measuring students’ learning of CT in class (S_CTCLS). The scale reflects the extent to which students report having learned various CT-related skills, such as using a solution for one problem to help solve another, breaking down complex problems into smaller steps, making diagrams to explain concepts, planning tasks in a structured manner, and detecting patterns in data.

Higher scores on this scale indicate that students perceive they have learned CT skills to a greater extent in their classes. The scale is composed of 9 items, item wording is presented in Table 12.26. The IRT results are very good (Table 12.27). Both outfit and infit are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.26: Item wording for the scale Learning about CT in class (S_CTCLS)

Scale or item	Question or item wording
S_CTCLS	To what extent have you learned how to do the following tasks when studying at school? <i>Recoded response categories for scaling: 0=Not at all;1=To a small extent;2=To a moderate extent;3=To a large extent</i>
IS3G26A	Use a solution that works for one problem to help solve a different problem
IS3G26B	Solve a hard problem by splitting it into a few easier problems
IS3G26C	Make diagrams that explain concepts or systems (e.g., electric circuits, plant growth, the water cycle)
IS3G26D	Plan tasks by making a list of the tasks in the order they need to be completed
IS3G26E	Detect patterns in data
IS3G26F	Use simulations to help understand concepts or systems (e.g., electric circuits, plant growth, growth of cities in a virtual world)
IS3G26G	Make flow diagrams to show how a computer program should work
IS3G26H	Systematically test computer programs to find bugs, errors, or other problems
IS3G26I	Use data to better understand real-world problems

Table 12.27: Item parameters, item fits, and WRMSD for the scale Learning about CT in class (S_CTCLS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IS3G26A	-0.652	-1.585	-0.640	2.225	0.994	1.025	0.067
IS3G26B	-0.375	-1.738	-0.461	2.199	0.983	1.014	0.064
IS3G26C	-0.116	-1.698	-0.289	1.987	1.030	1.058	0.070
IS3G26D	-0.350	-1.672	-0.396	2.067	1.015	1.055	0.065
IS3G26E	0.095	-1.581	-0.337	1.918	0.929	0.956	0.072
IS3G26F	0.339	-1.524	-0.390	1.914	0.900	0.926	0.063
IS3G26G	0.535	-1.313	-0.414	1.727	0.900	0.932	0.072
IS3G26H	0.578	-1.197	-0.504	1.701	0.923	0.951	0.071
IS3G26I	-0.054	-1.468	-0.396	1.863	0.950	0.982	0.063

Teacher Scales

Teachers' ICT self-efficacy (T_ICTEFF)

In Question 7, teachers were asked to indicate how well they could perform various school-related tasks using ICT. For each task, they selected from the following response options: "I can do this very well," "I can do this moderately well," "I haven't done this, but I could find out how," or "I do not think I could do this." The responses to these items were used to derive a scale reflecting teachers' ICT self-efficacy (T_ICTEFF). This scale includes items related to evaluating the quality of online teaching resources, contributing to discussion forums, creating presentations, conducting online transactions, and preparing ICT-integrated lessons.

Higher values on this scale indicate greater confidence in using ICT for teaching-related tasks. The scale is composed of 12 items, item wording is presented in Table 12.28. The IRT results are acceptable (Table 12.29). Both outfits and infits are mostly close to 1. WRMSD values for the majority of items are less than .1 and for all items are less than .2, suggesting a general good fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.28: Item wording for the scale Teachers' ICT self-efficacy (T_ICTEFF)

Scale or item	Question or item wording
T_ICTEFF	How well can you do these tasks using ICT? <i>Recoded response categories for scaling: 0=I do not think I could do this;1=I haven't done this, but I could find out how;2=I know how to do this moderately well;3=I know how to do this very well</i>
IT3G07B	Evaluate the quality of teaching resources on the internet
IT3G07C	Contribute to a discussion forum / user group on the Internet (e.g., a wiki or blog)
IT3G07D	Produce presentations (e.g., [Microsoft PowerPoint, Apple Keynote, Google Slides] or a similar program) with simple animation functions
IT3G07E	Use the Internet for online purchases and payments
IT3G07F	Prepare lessons that involve the use of ICT by students
IT3G07G	Use a spreadsheet program (e.g., [Microsoft Excel, Apple Numbers, Google Sheets]) for keeping records
IT3G07H	Use a spreadsheet program (e.g., Microsoft Excel, Apple Numbers, Google Sheets]) for analyzing data
IT3G07I	Assess student learning
IT3G07J	Collaborate with others using shared resources such as [Google Workspace, Office 365, Microsoft Teams, Zoho]
IT3G07L	Identify Internet scams
IT3G07M	Edit video content for use in teaching
IT3G07N	Create computer-based assessments that record students' responses to questions (e.g., [Kahoot, Google Forms, Microsoft Forms])

Table 12.29: Item parameters, item fits, and WRMSD for the scale Teachers' ICT self-efficacy (T_ICTEFF)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G07B	-0.704	-1.538	-0.725	2.263	1.034	1.066	0.077
IT3G07C	0.377	-2.400	0.465	1.935	1.264	1.242	0.105
IT3G07D	-0.457	-1.256	-0.481	1.737	0.815	0.880	0.079
IT3G07E	-0.930	-1.072	-0.243	1.315	1.065	1.103	0.095
IT3G07F	-0.507	-1.754	-0.499	2.253	0.825	0.865	0.083
IT3G07G	0.159	-1.742	-0.171	1.913	0.937	0.972	0.075
IT3G07H	0.523	-1.881	-0.015	1.896	0.986	1.008	0.084
IT3G07I	-0.487	-1.966	-0.271	2.237	0.918	0.972	0.093
IT3G07J	-0.062	-1.726	-0.244	1.969	0.990	1.016	0.104
IT3G07L	0.880	-1.609	-0.339	1.948	1.226	1.215	0.100
IT3G07M	0.898	-2.050	0.132	1.917	1.132	1.145	0.097
IT3G07N	0.309	-1.754	-0.020	1.775	0.935	0.959	0.078

Emphasis on ICT use in initial teacher education (T_EMPITE)

In Question 8, teachers were asked to indicate whether their initial teacher education (ITE) included training in various aspects of using ICT for teaching and learning. For each item, they selected from the following response options: "Yes," "No" or "I cannot remember." The responses to these items were used to derive a scale reflecting teachers' reports on the emphasis of ICT use in their initial teacher education (T_EMPITE). This scale includes items related to general ICT approaches for teaching, subject-specific ICT use, using ICT for teacher collaboration, using ICT for student assessment, and supporting students' ability to evaluate digital information.

Higher values on this scale indicate greater exposure to ICT training in initial teacher education programs. The scale is composed of 8 items, item wording is presented in Table 12.30. The IRT results are good (Table 12.31). Both outfit and infit appear smaller and are not close to 1 compared to other scales, which is because binary items have less variability. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.30: Item wording for the scale *Emphasis on ICT use in initial teacher education* (T_EMPITE)

Scale or item	Question or item wording
T_EMPITE	Did your [initial teacher education] include the following aspects of using ICT to support teaching and learning? <i>Recoded response categories for scaling: 0=No;1=Yes;Missing=I cannot remember</i>
IT3G08A	The use of productivity applications (e.g., word processor, presentation software, internet use, spreadsheets)
IT3G08B	General approaches (relevant across subjects) to using ICT to enhance teaching and learning
IT3G08C	Subject-specific approaches to using ICT to enhance teaching and learning
IT3G08D	Using ICT to collaborate with other teachers
IT3G08E	Using ICT to assess student learning
IT3G08F	Supporting students' capabilities to evaluate the reliability of Internet-based information sources
IT3G08G	Managing social problems that students experience when using ICT to communicate with others (e.g., cyberbullying)
IT3G08H	Supporting students' use of ICT as a tool for problem-solving

Table 12.31: Item parameters, item fits, and WRMSD for the scale *Emphasis on ICT use in initial teacher education* (T_EMPITE)

Item	Delta	Outfit	Infit	WRMSD
IT3G08A	-1.069	0.884	0.935	0.067
IT3G08B	-0.700	0.576	0.735	0.076
IT3G08C	-0.315	0.587	0.753	0.076
IT3G08D	0.058	0.599	0.762	0.061
IT3G08E	0.147	0.518	0.707	0.076
IT3G08F	0.629	0.607	0.769	0.072
IT3G08G	0.786	0.667	0.832	0.079
IT3G08H	0.464	0.493	0.701	0.072

Teacher participation in professional learning activities (T_PROFLRN)

In Question 9.i, teachers were asked to indicate how often they had participated in various professional learning activities related to ICT. For each activity, they selected from the following response options: "I did not participate in this activity," "Once only," or "More than once." The responses to these items were used to derive a scale reflecting teachers' participation in ICT-related professional learning (T_PROFLRN). This scale includes items related to training on subject-specific digital teaching applications, the use of productivity applications, supporting personalized student learning with ICT, managing students' social issues related to ICT, and helping students evaluate digital content. Higher values on this scale indicate greater engagement in professional learning related to ICT integration in teaching.

The scale is composed of 8 items, item wording is presented in Table 12.32. The IRT results are good (Table 12.33). Both outfit and infit are mostly close to 1. WRMSD values are all less than .1,

suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.32: Item wording for the scale Teacher participation in professional learning activities (T_PROFLRN)

Scale or item	Question or item wording
T_PROFLRN	How often have you participated in professional learning activities dealing with the following content areas in the past two years? <i>Recoded response categories for scaling: 0=I did not participate in this activity;1=Once only;2=More than once</i>
IT3G09AB	The use of subject-specific digital teaching and learning resources
IT3G09AA	The use of productivity applications (e.g., word processor, presentation software, internet use, spreadsheets) internet
IT3G09AD	How to use ICT to support students' personalized learning
IT3G09AE	Managing social problems that students experience when using ICT to communicate with others (e.g., cyberbullying)
IT3G09AF	Supporting students' capabilities to evaluate the reliability of Internet-based information sources
IT3G09AG	Supporting students' capabilities to identify deceptive Internet practices (e.g., scams, fake news, fake images, fake reviews, bots, etc.)
IT3G09AH	The use of visual coding platforms (e.g. [Alice, GameMaker, Kodu, Le.g.,o Mindstorms, MIT App Inventor, Scratch]) for teaching and learning
IT3G09AI	Integrating ICT into teaching and learning

Table 12.33: Item parameters, item fits, and WRMSD for the scale Teacher participation in professional learning activities (T_PROFLRN)

Item	Delta	Tau.Cat1	Tau.Cat2	Outfit	Infit	WRMSD
IT3G09AA	-0.465	-0.180	0.180	1.044	1.075	0.078
IT3G09AB	-0.962	-0.455	0.455	0.899	0.955	0.065
IT3G09AD	0.105	-0.178	0.178	0.897	0.972	0.082
IT3G09AE	0.189	-0.350	0.350	0.992	1.006	0.074
IT3G09AF	0.361	-0.210	0.210	0.673	0.796	0.069
IT3G09AG	0.578	-0.188	0.188	0.770	0.868	0.064
IT3G09AH	1.124	0.071	-0.071	1.325	1.195	0.067
IT3G09AI	-0.928	-0.404	0.404	0.860	0.929	0.073

Teacher perceptions of need for professional learning (T_PROFNEED)

In Question 9.ii, teachers were asked to indicate whether they believed they needed to do more professional learning related to various ICT-related topics. For each topic, they selected from the following response options: "I need to do more professional learning related to this content" or "I do not need to do more professional learning related to this content." The responses to these items were used to derive a scale reflecting teachers' perceived need for further professional learning in ICT (T_PROFNEED). This scale includes items related to the use of productivity applications, subject-specific digital teaching applications, using ICT to support personalized student learning, managing social problems related to ICT, and supporting students' ability to evaluate digital content.

Higher values on this scale indicate a greater perceived need for additional ICT-related professional learning. The scale is composed of 8 items, item wording is presented in Table 12.34. The IRT results are good (Table 12.35). Both outfits and infits appear smaller and are not close to 1 compared to other scales, which is because binary items have less variability. WRMSD values are all less than .1,



suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.34: Item wording for the scale *Teacher perceptions of need for professional learning* (T_PROFNEED)

Scale or item	Question or item wording
T_PROFNEED	Do you need to do more professional learning activities dealing with the following content areas in the future? <i>Recoded response categories for scaling: 0=I do not need to do more professional learning related to this content; 1=I need to do more professional learning related to this content</i>
IT3G09BA	The use of productivity applications (e.g., word processor, presentation software, internet use, spreadsheets) internet
IT3G09BB	The use of subject-specific digital teaching and learning resources
IT3G09BD	How to use ICT to support students' personalized learning
IT3G09BE	Managing social problems that students experience when using ICT to communicate with others (e.g., cyberbullying)
IT3G09BF	Supporting students' capabilities to evaluate the reliability of internet-based information sources
IT3G09BG	Supporting students' capabilities to identify deceptive internet practices (e.g., scams, fake news, fake images, fake reviews, bots, etc.)
IT3G09BH	The use of visual coding platforms (e.g. [Alice, GameMaker, Kodu, Lego, Mindstorms, MIT App Inventor, Scratch]) for teaching and learning
IT3G09BI	Integrating ICT into teaching and learning

Table 12.35: Item parameters, item fits, and WRMSD for the scale *Teacher perceptions of need for professional learning* (T_PROFNEED)

Item	Delta	Outfit	Infit	WRMSD
IT3G09BA	1.138	0.915	0.957	0.085
IT3G09BB	0.110	0.744	0.892	0.059
IT3G09BD	-0.308	0.688	0.855	0.056
IT3G09BE	-0.187	0.705	0.853	0.056
IT3G09BF	-0.015	0.595	0.779	0.062
IT3G09BG	-0.275	0.628	0.807	0.058
IT3G09BH	-0.547	1.031	1.059	0.076
IT3G09BI	0.084	0.711	0.865	0.058

Collaboration between teachers in using ICT (T_COLICT)

In Question 10, teachers were asked to indicate the extent to which they agreed with various statements regarding their collaboration with colleagues in using ICT for teaching. For each statement, they selected from the following response options: "Strongly agree," "Agree," "Disagree," or "Strongly disagree." The responses to these items were used to derive a scale reflecting teachers' collaboration in using ICT (T_COLICT). This scale includes items related to collaborating with other teachers to improve ICT use in teaching, developing ICT-based resources with colleagues, observing how others use ICT in teaching, discussing ICT use with colleagues, and sharing ICT-based resources.

Higher values on this scale indicate greater collaboration among teachers in integrating ICT into their instructional practices. The scale is composed of 8 items, item wording is presented in Table 12.36. The IRT results are good (Table 12.37). Both outfits and infits are all close to 1. WRMSD values are all less than .1 except for item IT3G10F (0.101), suggesting a good fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.36: Item wording for the scale Collaboration between teachers in using ICT (T_COLICT)

Scale or item	Question or item wording
T_COLICT	To what extent do you agree or disagree with the following statements about your use of ICT in teaching and learning at your school? <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IT3G10A	I collaborate with other teachers on improving the use of ICT in classroom teaching
IT3G10B	I collaborate with colleagues to develop ICT-based lessons
IT3G10C	I observe how other teachers use ICT in teaching
IT3G10D	I discuss with other teachers how to use ICT in teaching
IT3G10E	I share ICT-based resources with other teachers in my school
IT3G10F	I collaborate with other teachers to find personalized ICT-based resources for individual students
IT3G10G	I collaborate with other teachers to keep track of ICT competencies taught across subjects
IT3G10H	I collaborate with other teachers to create teaching materials that include the use of ICT in the classroom

Table 12.37: Item parameters, item fits, and WRMSD for the scale Collaboration between teachers in using ICT (T_COLICT)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G10A	-0.347	-2.462	-0.696	3.159	0.842	0.896	0.082
IT3G10B	0.151	-2.856	-0.416	3.272	0.814	0.842	0.093
IT3G10C	-0.295	-2.460	-0.905	3.365	1.134	1.173	0.080
IT3G10D	-0.626	-2.397	-1.099	3.496	0.919	0.977	0.086
IT3G10E	-0.377	-2.487	-0.746	3.233	0.931	1.008	0.096
IT3G10F	0.452	-2.926	-0.307	3.233	0.854	0.870	0.101
IT3G10G	0.647	-2.961	-0.282	3.243	0.931	0.937	0.097
IT3G10H	0.395	-2.751	-0.456	3.207	0.865	0.887	0.098

Availability of computer resources at school (T_RESRC)

In Question 13, teachers were asked to indicate the extent to which they agreed with various statements regarding the availability and functionality of ICT resources in their school. For each statement, they selected from the following response options: “Strongly agree,” “Agree,” “Disagree,” or “Strongly disagree.” The responses to these items were used to derive a scale reflecting teachers’ perceptions of ICT resource availability at school (T_RESRC). This scale includes items related to the adequacy of ICT equipment, the up-to-dateness of ICT resources, the reliability of ICT equipment, the sufficiency of technical support, and the quality of internet connectivity.

Higher values on this scale indicate more positive perceptions of ICT resource availability and functionality in schools. The scale is composed of 6 items, item wording is presented in Table 12.38. The IRT results are acceptable (Table 12.39). Both outfit and infit are mostly close to 1, WRMSD values are all less than .1 except for item IT3G13H (0.102), suggesting a good fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.38: Item wording for the scale Availability of computer resources at school (T_RESRC)

Scale or item	Question or item wording
T_RESRC	To what extent do you agree or disagree with the following statements about the use of ICT in teaching at your school? <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IT3G13A	My school has enough ICT equipment (e.g., computers and peripheral devices)
IT3G13B	The ICT equipment in my school is up to date
IT3G13C	The ICT equipment in my school works whenever I need to use it
IT3G13E	There is sufficient technical support to maintain ICT equipment
IT3G13F	My school has good connectivity (e.g., fast speed and stable) to the internet
IT3G13H	There is enough time to prepare lessons that incorporate ICT

Table 12.39: Item parameters, item fits, and WRMSD for the scale Availability of computer resources at school (T_RESRC)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G13A	-0.393	-2.071	-0.243	2.314	0.868	0.902	0.083
IT3G13B	-0.182	-2.240	-0.327	2.566	0.784	0.799	0.088
IT3G13C	-0.117	-2.535	-0.253	2.788	0.767	0.778	0.092
IT3G13E	0.003	-2.456	-0.282	2.738	0.915	0.927	0.074
IT3G13F	-0.030	-2.204	-0.389	2.593	1.081	1.097	0.071
IT3G13H	0.720	-2.586	-0.191	2.776	1.336	1.324	0.102

Negative views on using ICT in teaching and learning (T_VWNEG)

Two scales were constructed from the statements in Question 14. Teachers were asked to indicate the extent to which they agreed with various statements regarding potential negative consequences of ICT use in education. For each statement, they selected from the following response options: "Strongly agree," "Agree," "Disagree," or "Strongly disagree." The responses to these items were used to derive a scale reflecting teachers' negative views on using ICT for teaching and learning (T_VWNEG). This scale includes items related to concerns that ICT makes it difficult for students to develop deep thinking, leads to increased copying of material from the internet, distracts students from learning, results in poorer written expression, and negatively affects calculation and estimation skills.

Higher values on this scale indicate stronger concerns among teachers about the negative impacts of ICT use on student learning. The scale is composed of 8 items, item wording is presented in Table 12.40. The IRT results are acceptable (Table 12.41). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1 except for item IT3G14F (0.106), suggesting a general good fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.40: Item wording for the scale Negative views on using ICT in teaching and learning (T_VWNEG)

Scale or item	Question or item wording
T_VWNEG	To what extent do you agree or disagree with the following statements about using ICT in teaching and learning at your school? Using ICT at school ... <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IT3G14A	Makes it difficult for students to develop a deep understanding of concepts
IT3G14D	Results in students copying material from Internet sources without attribution
IT3G14F	Distracts students from learning
IT3G14G	Results in poorer written expression among students
IT3G14H	Results in poorer calculation and estimation skills among students
IT3G14I	Limits the amount of personal communication among students
IT3G14N	Results in shorter attention spans among students
IT3G14O	Confuses students with false or misleading information

Table 12.41: Item parameters, item fits, and WRMSD for the scale Negative views on using ICT in teaching and learning (T_VWNEG)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G14A	1.185	-3.164	0.690	2.473	1.250	1.247	0.071
IT3G14D	-0.673	-3.059	-0.073	3.133	1.204	1.207	0.074
IT3G14F	0.323	-3.214	0.571	2.643	0.942	0.947	0.106
IT3G14G	-0.503	-2.896	0.205	2.691	0.869	0.878	0.087
IT3G14H	-0.087	-3.302	0.497	2.805	0.866	0.874	0.083
IT3G14I	-0.176	-2.904	0.378	2.527	1.032	1.033	0.080
IT3G14N	-0.275	-3.153	0.298	2.855	0.963	0.963	0.083
IT3G14O	0.205	-3.299	0.234	3.064	0.987	0.991	0.063

Positive views on using ICT in teaching and learning (T_VWPOS)

Another scale was constructed from the statements in Question 14. Teachers were asked to indicate the extent to which they agreed with various statements regarding the potential benefits of ICT use in education. For each statement, they selected from the following response options: “Strongly agree,” “Agree,” “Disagree,” or “Strongly disagree.” The responses to these items were used to derive a scale reflecting teachers’ positive views on using ICT for teaching and learning (T_VWPOS). This scale includes items related to beliefs that ICT helps students develop a greater interest in learning, supports differentiated instruction by allowing students to work at an appropriate level, enhances problem-solving skills, facilitates student collaboration, and improves students’ ability to plan and organize their work.

Higher values on this scale indicate stronger positive perceptions of ICT’s role in teaching and learning. The scale is composed of 7 items, item wording is presented in Table 12.42. The IRT results are very good (Table 12.43). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.42: Item wording for the scale *Positive views on using ICT in teaching and learning (T_VWPOS)*

Scale or item	Question or item wording
T_VWPOS	To what extent do you agree or disagree with the following statements about using ICT in teaching and learning at your school? Using ICT at school ... <i>Recoded response categories for scaling: 0=Strongly disagree;1=Disagree;2=Agree;3=Strongly agree</i>
IT3G14B	Helps students develop a greater interest in learning
IT3G14C	Helps students to work at a level appropriate to their learning needs
IT3G14E	Helps students develop problem-solving skills
IT3G14J	Enables students to collaborate more effectively
IT3G14K	Helps students develop skills in planning and self-regulation of their work
IT3G14L	Improves academic performance of students
IT3G14M	Enables students to access better sources of information

Table 12.43: Item parameters, item fits, and WRMSD for the scale *Positive views on using ICT in teaching and learning (T_VWPOS)*

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G14B	-0.599	-2.588	-0.917	3.504	0.923	0.949	0.073
IT3G14C	-0.448	-3.123	-0.733	3.856	0.866	0.891	0.059
IT3G14E	0.246	-3.318	-0.439	3.758	0.938	0.946	0.073
IT3G14J	0.250	-3.333	-0.381	3.714	1.021	1.033	0.063
IT3G14K	0.554	-3.385	-0.328	3.712	0.912	0.920	0.070
IT3G14L	0.727	-3.436	-0.351	3.787	0.880	0.887	0.084
IT3G14M	-0.730	-2.621	-1.093	3.714	1.168	1.190	0.079

Teacher centered teaching approach (T_TEAPTC)

Three scales were constructed from the statements in Question 16.i. Teachers were asked to indicate how often students in their reference class engaged in various teacher-directed learning activities. For each activity, they selected from the following response options: "Never," "In some lessons," "In most lessons," or "In every or almost every lesson." The responses to these items were used to derive a scale reflecting teachers' use of teacher-centered teaching approaches with ICT (T_TEAPTC). This scale includes items related to presenting information to the whole class, conducting whole-class discussions, leading whole-class checking of schoolwork, having students work individually on assignments, and administering teacher-provided tests.

Higher values on this scale indicate more frequent use of teacher-directed instructional strategies in the classroom. The scale is composed of 5 items, item wording is presented in Table 12.44. The IRT results are good (Table 12.45). Both outfit and infit are all close to 1. WRMSD values are all less than .1 except for items IT3G16AB (0.106) and IT3G16AE (0.112), suggesting a general good fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.44: Item wording for the scale Teacher centered teaching approach (T_TEAPTC)

Scale or item	Question or item wording
T_TEAPTC	Think about your teaching of the reference class in this school year (i. How often did you and your students engage in the following activities?) <i>Recoded response categories for scaling: 0=Never;1=In some lessons;2=In most lessons;3=In every or almost every lesson;</i>
IT3G16AA	I presented information to the whole class
IT3G16AB	I conducted whole-class discussions
IT3G16AC	I led whole-class checking of schoolwork (e.g., going over solutions to worksheets or assignments, presenting answers questions on tests)
IT3G16AD	Students worked individually on worksheets, exercises, or assignments in class (with my help as required or requested during class)
IT3G16AE	Students completed tests provided by me

Table 12.45: Item parameters, item fits, and WRMSD for the scale Teacher centered teaching approach (T_TEAPTC)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G16AA	-1.067	-2.237	0.700	1.537	1.070	1.054	0.081
IT3G16AB	0.104	-2.120	0.583	1.538	1.139	1.144	0.106
IT3G16AC	0.180	-1.898	0.562	1.336	0.901	0.918	0.075
IT3G16AD	0.016	-2.195	0.579	1.616	0.959	0.966	0.077
IT3G16AE	0.766	-2.434	0.993	1.442	1.080	1.094	0.112

Inquiry-based teaching approach (T_TEAPIN)

Another scale was constructed from the statements in Question 16.i. Teachers were asked to indicate how often students in their reference class engaged in various ICT-supported inquiry-based learning activities. For each activity, they selected from the following response options: “Never,” “In some lessons,” “In most lessons,” or “In every or almost every lesson.” The responses to these items were used to derive a scale reflecting teachers’ use of inquiry-based teaching approaches with ICT (T_TEAPIN). This scale includes items related to helping students plan research or inquiries, guiding them in conducting academic or subject-related research, supporting students’ inquiry-based investigations, providing feedback on research activities, and facilitating student presentations of research findings using ICT.

Higher values on this scale indicate more frequent use of ICT to support inquiry-based learning strategies. The scale is composed of 5 items, item wording is presented in Table 12.46. The IRT results are good (Table 12.47). Both outfit and infit are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.46: Item wording for the scale Inquiry-based teaching approach (T_TEAPIN)

Scale or item	Question or item wording
T_TEAPIN	Think about your teaching of the reference class in this school year (i. How often did you and your students engage in the following activities?) <i>Recoded response categories for scaling: 0=Never;1=In some lessons;2=In most lessons;3=In every or almost every lesson;</i>
IT3G16AF	I helped students to plan research or inquiries (e.g., establishing research questions, setting constraints, etc) into academic/subject-specific topics
IT3G16AG	I helped students to conduct academic/subject-specific research by providing curriculum materials or instructions
IT3G16AH	I helped students to conduct research inquiries into real-world topics by asking critical/evaluative questions about their work
IT3G16AI	I gave feedback to students on their research work-in-progress
IT3G16AJ	Students presented the results of their research inquiries to the class

Table 12.47: Item parameters, item fits, and WRMSD for the scale Inquiry-based teaching approach (T_TEAPIN)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G16AF	0.266	-2.789	0.897	1.892	0.831	0.854	0.076
IT3G16AG	-0.038	-2.944	0.878	2.066	0.800	0.814	0.088
IT3G16AH	0.054	-2.746	0.727	2.019	0.774	0.799	0.082
IT3G16AI	-0.565	-2.663	0.699	1.964	0.893	0.906	0.077
IT3G16AJ	0.284	-3.475	1.287	2.188	1.076	1.088	0.068

Experiential/scenario-based teaching approach (T_TEAPEX)

Another scale was constructed from the statements in Question 16.i. Teachers were asked to indicate how often students in their reference class engaged in various ICT-supported experiential learning activities. For each activity, they selected from the following response options: “Never,” “In some lessons,” “In most lessons,” or “In every or almost every lesson.” The responses to these items were used to derive a scale reflecting teachers’ use of experiential and scenario-based teaching approaches with ICT (T_TEAPEX). This scale includes items related to helping students organize ideas to solve real-world problems, guiding students in searching for information in digital environments, and facilitating student collaboration with external individuals or groups using ICT.

Higher values on this scale indicate more frequent use of ICT to support experiential and scenario-based teaching strategies. The scale is composed of 3 items, item wording is presented in Table 12.48. The IRT results are very good (Table 12.49). Both outfits and infits are all close to 1, WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.



Table 12.48: Item wording for the scale Experiential/scenario-based teaching approach (T_TEAPEX)

Scale or item	Question or item wording
T_TEAPEX	Think about your teaching of the reference class in this school year (i. How often did you and your students engage in the following activities?) <i>Recoded response categories for scaling: 0=Never;1=In some lessons;2=In most lessons;3=In every or almost every lesson;</i>
IT3G16AK	I helped students to organize ideas to try to understand real-world problems
IT3G16AM	I guided students' searches for information into the real-world problems they were investigating
IT3G16AN	I helped students to work with people outside of the school to support the students' research into real-world problems

Table 12.49: Item parameters, item fits, and WRMSD for the scale Experiential/scenario-based teaching approach (T_TEAPEX)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G16AK	-0.562	-3.024	0.623	2.401	0.959	0.958	0.065
IT3G16AM	-0.418	-2.947	0.705	2.242	0.861	0.865	0.067
IT3G16AN	0.980	-1.790	0.476	1.314	0.920	0.987	0.077

Teachers' use of ICT for classroom activities - basic tasks (T_ICTCLASB)

Two scales were constructed from the statements in Question 17. Teachers were asked to indicate how often students in their reference class used ICT for various classroom activities. For each activity, they selected from the following response options: "Never," "One quarter or less (but more than never)," "About half of the lessons," "About three-quarters of the lessons," or "Almost every lesson." The responses to these items were used to derive a scale reflecting teachers' use of ICT for multimedia and collaborative classroom activities (T_ICTCLASB). This scale includes items related to recording audio or video (e.g., discussions, presentations), using chat or video apps for collaboration, editing digital images or videos, manually collecting and entering data, and creating or editing digital documents or presentations.

Higher values on this scale indicate more frequent use of ICT for multimedia and collaborative activities in the classroom. The scale is composed of 5 items, item wording is presented in Table 12.50. The IRT results are very good (Table 12.51). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.50: Item wording for the scale Teachers' use of ICT for classroom activities - basic tasks (T_ICTCLASB)

Scale or item	Question or item wording
T_ICTCLASB	In approximately what proportion of lessons in your reference class do students do the following ICT-based activities? <i>Recoded response categories for scaling: 0=Never;1=One quarter or less (but more than never);2=One half or less (but more than one quarter);3=More than one-half (but less than three quarters);4=Three quarters or more;</i>
IT3G17A	Record audio or video (e.g., discussions, presentations, performances, etc.)
IT3G17B	Use group chat, voice, or video apps to collaborate with other students on their schoolwork
IT3G17C	Edit digital images, photos, or videos
IT3G17E	Collect and manually enter data
IT3G17M	Create or edit documents or presentations



Table 12.51: Item parameters, item fits, and WRMSD for the scale Teachers' use of ICT for classroom activities - basic tasks (T_ICTCLASB)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	Outfit	Infit	WRMSD
IT3G17A	0.319	-2.005	0.278	0.639	1.088	1.010	1.036	0.062
IT3G17B	0.263	-1.516	0.118	0.484	0.913	0.965	1.021	0.053
IT3G17C	0.125	-1.894	0.140	0.605	1.150	0.853	0.890	0.060
IT3G17E	0.124	-1.839	0.066	0.642	1.131	1.108	1.147	0.056
IT3G17M	-0.831	-2.144	0.066	0.791	1.287	1.181	1.187	0.065

Teachers' use of ICT for classroom activities - advanced tasks (T_ICTCLASA)

Another scale was constructed from the statements in Question 17. Teachers were asked to indicate how often students in their reference class used ICT for different classroom activities. For each activity, they selected from the following response options: "Never," "One quarter or less (but more than never)," "About half of the lessons," "About three-quarters of the lessons," or "Almost every lesson." The responses to these items were used to derive a scale reflecting teachers' use of ICT for classroom activities (T_ICTCLASA). This scale includes items related to using simulation software to understand concepts, recording sensor data from digital devices, creating digital charts from stored data, editing extended videos, and writing computer programs or apps.

Higher values on this scale indicate more frequent use of ICT for these activities in the classroom. The scale is composed of 6 items, item wording is presented in Table 12.52. The IRT results are good (Table 12.53). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.52: Item wording for the scale Teachers' use of ICT for classroom activities - advanced tasks (T_ICTCLASA)

Scale or item	Question or item wording
T_ICTCLASA	In approximately what proportion of lessons in your reference class do students do the following ICT-based activities? <i>Recoded response categories for scaling: 0=Never;1=One quarter or less (but more than never);2=One half or less (but more than one quarter);3=More than one-half (but less than three quarters);4=Three quarters or more;</i>
IT3G17D	Use simulation software to understand concepts or systems (e.g., electric circuits, plant growth, growth of cities in a virtual world, etc.)
IT3G17F	Record sensor data from digital devices (e.g., the accelerometer in a smartphone or robot)
IT3G17G	Create digital charts from stored data
IT3G17I	Create or edit extended (more than 5 minute) videos or animations for a specific purpose and audience
IT3G17J	Write computer programs, games, or apps (e.g., using [Python, LUA, Javascript, Scratch])
IT3G17K	Create digital flowcharts or decision trees to illustrate complex systems



Table 12.53: Item parameters, item fits, and WRMSD for the scale Teachers' use of ICT for classroom activities - advanced tasks (T_ICTCLASA)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	Outfit	Infit	WRMSD
IT3G17D	-0.248	-1.577	-0.208	0.559	1.226	1.127	1.165	0.046
IT3G17F	0.044	-1.271	-0.212	0.389	1.094	0.728	0.857	0.059
IT3G17G	-0.122	-1.724	-0.166	0.625	1.265	0.863	0.907	0.054
IT3G17I	-0.203	-1.799	-0.040	0.577	1.262	1.162	1.122	0.053
IT3G17J	0.201	-0.883	-0.339	0.266	0.957	0.799	0.953	0.052
IT3G17K	0.329	-1.054	-0.315	0.351	1.018	0.701	0.815	0.053

Use of digital learning tools (T_USETOOL)

Two scales were constructed from the statements in Question 18. Teachers were asked to indicate how often they used various ICT tools in their teaching of the reference class. For each tool, they selected from the following response options: "Never," "In some lessons," "In most lessons," or "In every or almost every lesson." The responses to these items were used to derive a scale reflecting teachers' use of digital learning tools (T_USETOOL). This scale includes items related to digital learning games, concept mapping software, simulations and modeling tools, graphing or drawing software, and social media for educational purposes.

Higher values on this scale indicate more frequent use of these digital learning tools in classroom instruction. The scale is composed of 7 items, item wording is presented in Table 12.54. The IRT results are very good (Table 12.55). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.54: Item wording for the scale Use of digital learning tools (T_USETOOL)

Scale or item	Question or item wording
T_USETOOL	How often did you use the following tools in your teaching of the reference class this school year? Recoded response categories for scaling: 0=Never;1=In some lessons;2=In most lessons;3=In every or almost every lesson
IT3G18B	Digital learning games
IT3G18G	Concept mapping software (e.g., [Inspiration, Webspiration])
IT3G18H	Simulations and modelling software (e.g., [NetLogo])
IT3G18J	Graphing or drawing software
IT3G18M	Social media (e.g., [Instagram, SnapChat, Twitter])
IT3G18O	Virtual or augmented reality (e.g., [The Body VR, Google Earth VR, Math Alive])
IT3G18P	Adaptive learning systems (software that gathers and uses student data to deliver personalized resources and learning activities to address the unique needs of students)



Table 12.55: Item parameters, item fits, and WRMSD for the scale Use of digital learning tools (T_USETOOL)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G18B	-0.826	-2.585	0.650	1.934	1.119	1.117	0.068
IT3G18G	0.320	-1.514	0.310	1.204	0.811	0.853	0.068
IT3G18H	0.508	-0.977	0.041	0.936	0.904	0.890	0.057
IT3G18J	-0.166	-1.604	0.297	1.306	1.085	1.042	0.057
IT3G18M	-0.390	-1.455	0.357	1.098	1.082	1.084	0.078
IT3G18O	0.343	-1.295	0.314	0.981	0.866	0.870	0.059
IT3G18P	0.211	-1.272	0.251	1.021	0.861	0.897	0.064

Use of general utility software (T_USEUTIL)

Another scale was constructed from the statements in Question 18. Teachers were asked to indicate how often they used various ICT tools in their teaching of the reference class. For each tool, they selected from the following response options: “Never,” “In some lessons,” “In most lessons,” or “In every or almost every lesson.” The responses to these items were used to derive a scale reflecting teachers’ use of general utility software (T_USEUTIL). This scale includes items related to word-processing software, presentation software, spreadsheets, video and photo editing software, and computer-based information resources such as websites and online encyclopedias.

Higher values on this scale indicate more frequent use of these general utility tools in classroom instruction. The scale is composed of 6 items, item wording is presented in Table 12.56. The IRT results are good (Table 12.57). Both outfits and infits are mostly close to 1, WRMSD values are all less than .1 except for item IT3G18L (0.108), suggesting a general good fit. RMSD values for all items and all countries are below .25 except for item IT3G18L for Malta (0.259), indicating that this scale is invariant across countries except for Malta.

Table 12.56: Item wording for the scale Use of general utility software (T_USEUTIL)

Scale or item	Question or item wording
T_USEUTIL	How often did you use the following tools in your teaching of the reference class this school year? <i>Recoded response categories for scaling: 0=Never;1=In some lessons;2=In most lessons;3=In every or almost every lesson</i>
IT3G18C	Word-processor software (e.g., [Microsoft Word, Apple Pages, Google Docs])
IT3G18D	Presentation software (e.g., [Microsoft PowerPoint, Apple Keynote, Google Slides])
IT3G18E	Spreadsheets (e.g., [Microsoft Excel, Apple Numbers, Google Sheets])
IT3G18F	Video and photo software for capture and editing (e.g., [Windows Movie Maker, iMovie, Adobe Photoshop])
IT3G18I	Computer-based information resources (e.g., websites, wikis, encyclopedia)
IT3G18L	Digital content linked with paper-based textbooks



Table 12.57: Item parameters, item fits, and WRMSD for the scale Use of general utility software (T_USEUTIL)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G18C	-0.522	-2.095	0.502	1.593	0.927	0.939	0.081
IT3G18D	-0.665	-2.093	0.697	1.396	0.951	0.954	0.085
IT3G18E	0.829	-1.584	0.500	1.084	1.003	1.033	0.083
IT3G18F	1.158	-1.585	0.527	1.059	1.135	1.116	0.088
IT3G18I	-0.369	-2.219	0.519	1.700	1.009	1.030	0.075
IT3G18L	-0.431	-1.406	0.400	1.006	1.315	1.314	0.108

Emphasis on ICT capabilities in class (T_ICTEMP)

In Question 19, teachers were asked to indicate the emphasis they had given to developing different ICT-based skills among students in their reference class. For each skill, they selected from the following response options: “Strong emphasis,” “Some emphasis,” “Little emphasis,” or “No emphasis.” The responses to these items were used to derive a scale reflecting teachers’ emphasis on developing ICT capabilities in class (T_ICTEMP). This scale includes items related to displaying information for a specific audience, evaluating the credibility of digital information, sharing digital content, using software to construct digital representations, and providing digital feedback on others’ work.

Higher values on this scale indicate greater emphasis on developing students’ ICT-related skills in the classroom. The scale is composed of 12 items, item wording is presented in Table 12.58. The IRT results are good (Table 12.59). Both outfits and infits are mostly close to 1, WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.58: Item wording for the scale Emphasis on ICT capabilities in class (T_ICTEMP)

Scale or item	Question or item wording
T_ICTEMP	In your teaching of the reference class in this school year, how much emphasis have you given to developing the following ICT-based capabilities in your students? <i>Recoded response categories for scaling: 0=No emphasis;1=Little emphasis;2=Some emphasis;3=Strong emphasis</i>
IT3G19B	To display information for a given audience/purpose
IT3G19C	To evaluate the credibility of digital information
IT3G19D	To share digital information with others
IT3G19E	To use computer software to construct digital work products (e.g., presentations, documents, images and diagrams, etc.)
IT3G19F	To provide digital feedback on the work of others (such as their classmates)
IT3G19G	To explore a range of digital resources when searching for information
IT3G19H	To provide references for digital information sources
IT3G19I	To understand the consequences of making information publicly available online
IT3G19J	To collaborate with their classmates using an online collaboration platform (e.g., [Google Workspace, Office 365, Microsoft Teams, Zoho])
IT3G19K	To refine internet searches to return fewer or more relevant results
IT3G19M	To identify deceptive internet practices (e.g., scams, fake news, fake images, fake reviews, bots, etc.)
IT3G19N	To check if facts from internet-based sources are consistent with other sources



Table 12.59: Item parameters, item fits, and WRMSD for the scale *Emphasis on ICT capabilities in class* (T_ICTEMP)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G19B	-0.551	-1.943	-0.514	2.456	1.134	1.162	0.087
IT3G19C	-0.512	-1.876	-0.346	2.222	0.933	0.961	0.083
IT3G19D	-0.202	-2.028	-0.430	2.457	0.963	0.985	0.071
IT3G19E	-0.316	-1.686	-0.476	2.161	1.092	1.126	0.073
IT3G19F	0.680	-1.816	-0.314	2.130	1.032	1.068	0.091
IT3G19G	-0.279	-1.937	-0.443	2.380	0.862	0.893	0.081
IT3G19H	-0.045	-1.769	-0.340	2.109	0.900	0.941	0.088
IT3G19I	-0.204	-1.592	-0.307	1.899	0.964	1.012	0.080
IT3G19J	0.550	-1.694	-0.247	1.941	1.273	1.279	0.084
IT3G19K	0.342	-1.793	-0.336	2.129	0.900	0.927	0.079
IT3G19M	0.443	-1.489	-0.224	1.713	1.025	1.065	0.085
IT3G19N	0.095	-1.662	-0.322	1.984	0.886	0.930	0.073

Teacher emphasis on CT related tasks in class (T_CODEMP)

In Question 20, teachers were asked to indicate the emphasis they had given to developing different CT skills among students in their reference class. For each skill, they selected from the following response options: “Strong emphasis,” “Some emphasis,” “Little emphasis,” or “No emphasis.” The responses to these items were used to derive a scale reflecting teachers’ emphasis on CT skills in class (T_CODEMP). This scale includes items related to the emphasis placed on breaking down complex problems into smaller steps, creating diagrams to explain systems, detecting patterns in data, analyzing data for real-world insights, and describing the rules governing systems.

Higher values on this scale indicate greater emphasis on developing CT skills through classroom instruction. The scale is composed of 6 items, item wording is presented in Table 12.60. The IRT results are good (Table 12.61). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.60: Item wording for the scale *Teacher emphasis on CT related tasks in class* (T_CODEMP)

Scale or item	Question or item wording
T_CODEMP	In your teaching of the reference class this school year, how much emphasis have you given to teaching the following skills? <i>Recoded response categories for scaling: 0=No emphasis;1=Little emphasis;2=Some emphasis;3=Strong emphasis</i>
IT3G20B	To solve complex problems by splitting them into smaller problems
IT3G20C	To make diagrams that explain concepts or systems (e.g., electric circuits, plant growth, the water cycle, etc.)
IT3G20E	To detect patterns in data
IT3G20H	To analyze data to better understand real-world problems
IT3G20I	To describe the rules that govern how a system works (e.g., a vending machine, the school canteen, a game, etc.)
IT3G20J	To evaluate and improve solutions to real-world problems



Table 12.61: Item parameters, item fits, and WRMSD for the scale Teacher emphasis on CT related tasks in class (T_CODEMP)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit	WRMSD
IT3G20B	-0.817	-1.700	-0.556	2.256	1.029	1.068	0.073
IT3G20C	0.458	-1.288	-0.434	1.723	1.132	1.141	0.072
IT3G20E	0.114	-1.371	-0.478	1.849	0.937	0.974	0.084
IT3G20H	-0.049	-1.425	-0.490	1.915	0.798	0.838	0.078
IT3G20I	0.564	-1.343	-0.376	1.718	0.956	0.991	0.070
IT3G20J	-0.269	-1.583	-0.462	2.045	0.882	0.914	0.078

Teacher epistemological beliefs-Embodied cognition (T_EPIEMC)

Three scales were constructed from the statements in Question 21. Teachers were asked to indicate the extent to which they agreed with various statements about knowledge and learning. For each statement, they selected from the following response options: “Not at all,” “To a very small extent,” “To a small extent,” “To a moderate extent,” “To a large extent,” or “Completely.” The responses to these items were used to derive a scale reflecting teachers’ epistemological beliefs aligned with embodied cognition (T_EPIEMC). This scale includes items related to understanding abstract phenomena by relating them to real-world experiences, the role of whole-body experiences in cognition, learning as an active and social process, and the idea that deep understanding emerges from unsystematic exploration.

Higher values on this scale indicate stronger alignment with embodied cognition perspectives on learning and knowledge development. The scale is composed of 5 items, item wording is presented in Table 12.62. The IRT results are good (Table 12.63). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.62: Item wording for the scale Teacher epistemological beliefs-Embodied cognition (T_EPIEMC)

Scale or item	Question or item wording
T_EPIEMC	To what extent does each of the following statements reflect your beliefs about knowledge, learning, and cognition? <i>Recoded response categories for scaling: 0=Not at all or to a very small extent;1=To a small extent;2=To a moderate extent;3=To a large extent;4=Completely</i>
IT3G21B	People understand abstract phenomena by relating them to physical experiences
IT3G21C	Learners construct their own understanding by internalizing their experiences
IT3G21E	Cognition depends on an individual's whole-body response to the physical world
IT3G21H	Knowledge is best created and maintained through personal action and reflection
IT3G21P	Deep understanding of concepts happens unsystematically and depends on the experiences made available to the learner



Table 12.63: Item parameters, item fits, and WRMSD for the scale Teacher epistemological beliefs-Embodied cognition (T_EPIEMC)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	Outfit	Infit	WRMSD
IT3G21B	-0.021	-1.440	-1.149	0.261	2.328	0.886	0.896	0.064
IT3G21C	-0.216	-1.344	-1.126	0.223	2.247	0.876	0.896	0.072
IT3G21E	0.296	-1.620	-1.197	0.458	2.358	1.027	1.025	0.065
IT3G21H	-0.500	-1.245	-1.181	0.114	2.313	0.991	0.997	0.069
IT3G21P	0.441	-1.632	-1.140	0.425	2.346	1.311	1.299	0.068

Teacher epistemological beliefs-Cognitivism (T_EPICOG)

Another scale was constructed from the statements in Question 21. Teachers were asked to indicate the extent to which they agreed with various statements about knowledge and learning. For each statement, they selected from the following response options: “Not at all,” “To a very small extent,” “To a small extent,” “To a moderate extent,” “To a large extent,” or “Completely.” The responses to these items were used to derive a scale reflecting teachers’ epistemological beliefs aligned with cognitivism (T_EPICOG). This scale includes items related to systematic engagement in knowledge construction, conceptual knowledge being based on fundamental truths, learning complex concepts through mastery of simpler ones, the importance of guidance before independent learning, and the necessity of maintaining knowledge through repeated practice.

Higher values on this scale indicate stronger alignment with cognitivist perspectives on learning and knowledge development. The scale is composed of 5 items, item wording is presented in Table 12.64. The IRT results are very good (Table 12.65). Both outfits and infits are all close to 1. WRMSD values are all less than .1, suggesting an excellent fit. RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.64: Item wording for the scale Teacher epistemological beliefs-Cognitivism (T_EPICOG)

Scale or item	Question or item wording
T_EPICOG	To what extent does each of the following statements reflect your beliefs about knowledge, learning, and cognition? <i>Recoded response categories for scaling: 0=Not at all or to a very small extent;1=To a small extent;2=To a moderate extent;3=To a large extent;4=Completely</i>
IT3G21A	Knowledge is developed through systematic engagement with basic facts that leads to deeper understanding in the future
IT3G21D	Conceptual knowledge is based on fundamental truths and the relationships between these truths
IT3G21F	Complex concepts are mastered by first mastering the underlying basic concepts
IT3G21I	The best learning happens when preceded by a general overview of the learning content
IT3G21L	Knowledge dissipates if it is not maintained through repetition and practice



Table 12.65: Item parameters, item fits, and WRMSD for the scale Teacher epistemological beliefs-Cognitivism (T_EPICOG)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	Outfit	Infit	WRMSD
IT3G21A	0.200	-1.479	-1.266	0.283	2.462	0.958	0.960	0.065
IT3G21D	0.344	-1.703	-1.434	0.460	2.677	1.002	0.998	0.068
IT3G21F	-0.258	-1.171	-1.067	0.277	1.961	0.916	0.940	0.069
IT3G21I	0.092	-1.908	-1.249	0.550	2.608	1.078	1.085	0.059
IT3G21L	-0.379	-1.415	-1.000	0.388	2.027	1.133	1.146	0.067

Teacher epistemological beliefs-Constructivist (T_EPICON)

Another scale was constructed from the statements in Question 21. Teachers were asked to indicate the extent to which they agreed with various statements about knowledge and learning. For each statement, they selected from the following response options: "Not at all," "To a very small extent," "To a small extent," "To a moderate extent," "To a large extent," or "Completely." The responses to these items were used to derive a scale reflecting teachers' epistemological beliefs aligned with constructivism (T_EPICON). This scale includes items related to learning as a social process that requires interaction, the importance of critically evaluating new information, the influence of prior experiences on learning quality, and the effectiveness of learning through hands-on exploration and play.

Higher values on this scale indicate stronger alignment with constructivist perspectives on learning and knowledge development. The scale is composed of 4 items, item wording is presented in Table 12.66. The IRT results are good (Table 12.67). Both outfits and infits are mostly close to 1. WRMSD values are all less than .1, suggesting an excellent fit, and RMSD values for all items and all countries are all below .25, indicating that this scale is invariant across countries.

Table 12.66: Item wording for the scale Teacher epistemological beliefs-Constructivist (T_EPICON)

Scale or item	Question or item wording
T_EPICON	To what extent does each of the following statements reflect your beliefs about knowledge, learning, and cognition? <i>Recoded response categories for scaling: 0=Not at all or to a very small extent;1=To a small extent;2=To a moderate extent;3=To a large extent;4=Completely</i>
IT3G21K	Learning is a social process requiring interaction with other people
IT3G21M	All new information should be critically evaluated
IT3G21N	The quality of learning depends on the experiences of the learner and the learning context
IT3G21O	The best learning happens when concepts are placed in their wider context

Table 12.67: Item parameters, item fits, and WRMSD for the scale Teacher epistemological beliefs-Constructivist (T_EPICON)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Tau.Cat4	Outfit	Infit	WRMSD
IT3G21K	0.049	-1.747	-1.195	0.528	2.414	1.198	1.206	0.055
IT3G21M	0.099	-1.470	-1.266	0.492	2.244	1.071	1.072	0.059
IT3G21N	-0.144	-1.779	-1.325	0.319	2.785	0.842	0.849	0.069
IT3G21O	-0.004	-1.974	-1.350	0.603	2.721	0.897	0.901	0.065

Principal Scales

Principals' views on using ICT for educational outcomes (P_VWICT)



In Question 11, school principals were asked to indicate how important they believed various ICT-related outcomes were for education in their school. For each outcome, they selected from the following response options: “Very important,” “Quite important,” “Somewhat important,” or “Not important.” The responses to these items were used to derive a scale reflecting principals’ views on using ICT for educational outcomes (P_VWICT). This scale includes items related to the perceived importance of developing students’ basic computer skills, enhancing students’ ability to use ICT for communication, fostering responsible ICT use, improving learning through ICT, and promoting students’ critical understanding of digital information.

Higher values on this scale indicate greater importance assigned by school principals to ICT-related skills as an outcome of education. The scale is composed of 10 items, item wording is presented in Table 12.68. The IRT results are good (Table 12.69). Infits are mostly close to 1, suggesting a general good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.68: Item wording for the scale Principals’ views on using ICT for educational outcomes (P_VWICT)

Scale or item	Question or item wording
P_VWICT	How important is each of the following outcomes of education in your school? <i>Recoded response categories for scaling: 0=Not important;1=Somewhat important;2=Quite important;3=Very important</i>
IP3G11A	The development of students’ basic computer skills (e.g., internet use, email, word processing, presentation software, etc.)
IP3G11B	The development of students’ skills in using ICT for collaboration
IP3G11C	The use of ICT for facilitating students’ responsibility for their own learning
IP3G11D	The use of ICT to augment and improve students’ learning
IP3G11E	The development of students’ understanding and skills relating to safe and appropriate use of ICT
IP3G11F	The development of students’ proficiency in accessing and using information with ICT
IP3G11G	The development of students’ capacity to use ICT to produce multimedia and digital content to communicate effectively with an audience
IP3G11H	The development of students’ ability to write or create apps or programs
IP3G11I	The development of students’ proficiency in protecting themselves from deceptive internet practices (e.g., scams, fake news, fake images, fake reviews, bots, etc.)
IP3G11J	The development of students’ proficiency in computational thinking

Table 12.69: Item parameters, item fits, and WRMSD for the scale Principals’ views on using ICT for educational outcomes (P_VWICT)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IP3G11A	-1.124	-1.935	-0.556	2.492	0.937	1.015
IP3G11B	-0.154	-2.071	-0.445	2.516	0.754	0.823
IP3G11C	0.237	-2.015	-0.322	2.337	0.828	0.880
IP3G11D	-0.223	-2.567	-0.100	2.667	0.741	0.826
IP3G11E	-0.923	-1.723	-0.769	2.492	0.700	0.881
IP3G11F	-0.830	-2.264	-0.492	2.756	0.684	0.814
IP3G11G	0.501	-2.822	0.031	2.790	0.894	0.962
IP3G11H	1.984	-2.062	-0.016	2.078	1.050	1.099
IP3G11I	-0.384	-1.235	-0.514	1.748	1.074	1.062
IP3G11J	0.916	-2.166	-0.042	2.207	0.944	1.027

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' reports on expectations of ICT use by teachers (P_EXPLRN)

Two scales were constructed from Question 12. School principals were asked to indicate whether teachers in their school were expected to acquire knowledge and skills in various ICT-related activities. For each activity, they selected from the following response options: "Expected and required," "Expected but not required," or "Not expected." The responses to these items were used to derive a scale reflecting principals' reports on expectations of ICT use by teachers (P_EXPLRN). This scale includes items related to expectations that teachers will integrate web-based learning into instruction, use ICT for student assessment and monitoring progress, incorporate subject-specific digital learning resources, and integrate ICT into teaching and learning.

Higher values on this scale indicate greater expectations among school principals regarding teachers' use of ICT in their instructional practices. The scale is composed of 10 items, item wording is presented in Table 12.70. The IRT results are good (Table 12.71). Both outfits and infits are mostly close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.70: Item wording for the scale Principals' reports on expectations of ICT use by teachers (P_EXPLRN)

Scale or item	Question or item wording
P_EXPLRN	Are teachers in your school expected to acquire knowledge and skills in each of the following activities? <i>Recoded response categories for scaling: 0=Not expected;1=Expected but not required;2=Expected and required</i>
IP3G12A	Integrate Web-based learning into their instructional practice
IP3G12B	Use ICT-based forms of student assessment
IP3G12C	Use ICT for monitoring student progress
IP3G12G	Integrate ICT into teaching and learning
IP3G12H	Use subject-specific digital learning resources (e.g., tutorials, simulation, etc.)
IP3G12I	Use e-portfolios for assessment
IP3G12J	Use ICT to develop authentic (real-life) assignments for students
IP3G12K	Assess students' [computer and information literacy]
IP3G12L	Use ICT to support [students with special needs or specific learning difficulties]
IP3G12N	Assess students' [computational thinking]

Table 12.71: Item parameters, item fits, and WRMSD for the scale Principals' reports on expectations of ICT use by teachers (P_EXPLRN)

Item	Delta	Tau.Cat1	Tau.Cat2	Outfit	Infit
IP3G12A	-0.395	-1.810	1.810	1.035	1.060
IP3G12B	0.193	-1.602	1.602	0.860	0.901
IP3G12C	0.072	-1.443	1.443	0.847	0.891
IP3G12G	-1.851	-2.323	2.323	0.861	0.961
IP3G12H	-0.595	-2.139	2.139	0.839	0.905
IP3G12I	1.381	-1.304	1.304	0.888	0.910
IP3G12J	0.750	-1.915	1.915	0.835	0.858
IP3G12K	-0.045	-1.257	1.257	0.940	1.006
IP3G12L	-0.465	-1.730	1.730	0.949	0.994
IP3G12N	0.956	-1.396	1.396	0.976	0.994

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' reports on expectations for teacher collaboration using ICT (P_EXPTCH)

Another scale was constructed from the statements in Question 12. School principals were asked to indicate whether teachers in their school were expected to acquire knowledge and skills in various ICT-related activities. For each activity, they selected from the following response options: "Expected and required," "Expected but not required," or "Not expected." The responses to these items were used to derive a scale reflecting principals' reports on expectations for teacher collaboration using ICT (P_EXPTCH). This scale includes items related to expectations that teachers will collaborate with other teachers via ICT, communicate with parents via ICT, and communicate with students via ICT.

Higher values on this scale indicate greater expectations among school principals regarding teachers' use of ICT for collaboration and communication. The scale is composed of 3 items, item wording is presented in Table 12.72. The IRT results are acceptable (Table 12.73). Both outfit and infit are mostly close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.72: Item wording for the scale Principals' reports on expectations for teacher collaboration using ICT (P_EXPTCH)

Scale or item	Question or item wording
P_EXPTCH	Are teachers in your school expected to acquire knowledge and skills in each of the following activities? <i>Recoded response categories for scaling: 0=Not expected;1=Expected but not required;2=Expected and required</i>
IP3G12D	Collaborate with other teachers via ICT
IP3G12E	Communicate with parents via ICT
IP3G12F	Communicate with students via ICT

Table 12.73: Item parameters, item fits, and WRMSD for the scale Principals' reports on expectations for teacher collaboration using ICT (P_EXPTCH)

Item	Delta	Tau.Cat1	Tau.Cat2	Outfit	Infit
IP3G12D	0.066	-2.137	2.137	0.970	1.029
IP3G12E	-0.206	-0.580	0.580	0.867	0.768
IP3G12F	0.140	-1.150	1.150	0.714	0.722

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' reports on priorities for facilitating use of ICT (P_PRIORS)

In Question 20, school principals were asked to indicate the priority given to different strategies for facilitating the use of ICT in teaching and learning. For each strategy, they selected from the following response options: "The school has no influence over this way of facilitating the use of ICT in teaching and learning," "High priority," "Medium priority," "Low priority," or "Not a priority." The responses to these items were used to derive a scale reflecting principals' reports on priorities for facilitating ICT use in teaching and learning (P_PRIORS). This scale includes items related to increasing the variety of digital learning resources, enhancing online learning systems, supporting professional development in ICT, expanding technical support availability, and providing incentives for teachers to integrate ICT into instruction.

Higher values on this scale indicate higher levels of priority assigned to these ICT support and implementation strategies in schools. The scale is composed of 10 items, item wording is presented in Table 12.74. The IRT results are good (Table 12.75). Both outfit and infit are mostly close to 1, suggesting



a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.74: Item wording for the scale Principals' reports on priorities for facilitating use of ICT (P_PRIORS)

Scale or item	Question or item wording
P_PRIORS	At your school, what priority is given to the following ways of facilitating the use of ICT in teaching and learning? <i>Recoded response categories for scaling: 0=Not a priority;1=Low priority;2=Medium priority;3=High priority;Missing=The school has no influence over this way of facilitating the use of ICT in teaching and learning</i>
IP3G20C	Increasing the variety of digital learning resources available for teaching and learning
IP3G20D	Establishing or enhancing an online learning support platform
IP3G20E	Supporting participation in professional development on the use of ICT in teaching and learning
IP3G20F	Increasing the availability of qualified technical personnel to support the use of ICT
IP3G20G	Providing teachers with incentives to integrate ICT use in their teaching
IP3G20H	Providing more time for teachers to prepare lessons in which ICT is used
IP3G20I	Increasing the professional learning resources for teachers in the use of ICT
IP3G20J	Fostering collaboration between teachers within the school to support the integration of ICT use in their teaching
IP3G20K	Fostering collaboration between teachers in this school and with teachers in other schools (e.g., teacher networks) to support the integration of ICT use in their teaching
IP3G20L	Developing a shared vision for using ICT to support teaching and learning

Table 12.75: Item parameters, item fits, and WRMSD for the scale Principals' reports on priorities for facilitating use of ICT (P_PRIORS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IP3G20C	-0.554	-0.749	-0.973	1.722	0.938	1.021
IP3G20D	0.398	-0.934	-0.512	1.446	1.215	1.230
IP3G20E	-0.838	-1.334	-0.621	1.955	0.847	0.944
IP3G20F	-0.054	-0.912	-0.364	1.276	1.177	1.123
IP3G20G	-0.150	-0.810	-0.670	1.479	1.166	1.219
IP3G20H	1.032	-1.021	-0.536	1.557	0.963	0.996
IP3G20I	-0.134	-1.309	-0.454	1.764	0.726	0.803
IP3G20J	-0.599	-1.265	-0.528	1.794	0.725	0.844
IP3G20K	0.941	-1.343	-0.355	1.698	1.024	1.062
IP3G20L	-0.041	-1.266	-0.352	1.618	0.803	0.896

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' use of ICT for general school-related activities (P_ICTUSE)

Two scales were constructed from Question 24. School principals were asked to indicate how often they used ICT for various general school-related activities. For each activity, they selected from the following response options: "Never," "Less than once a month," "At least once a month but not every week," "At least once a week but not every day," or "Every day." The responses to these items were used to derive a scale reflecting principals' use of ICT for general school-related activities (P_ICTUSE). This scale includes items related to the frequency with which school principals provide information on educational issues, maintain and analyze data using ICT, prepare presentations, work with learning management systems, and manage staff through ICT-based tools.

Higher values on this scale indicate more frequent use of ICT for general administrative and managerial tasks in school leadership. The scale is composed of 7 items, item wording is presented in Table 12.76. The IRT results are acceptable (Table 12.77). Both outfits and infits are mostly close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.76: Item wording for the scale Principals' use of ICT for general school-related activities (P_ICTUSE)

Scale or item	Question or item wording
P_ICTUSE	How often do you use ICT for the following activities? <i>Recoded response categories for scaling: 0=Never;0=Less than once a month;1=At least once a month but not every week;2=At least once a week but not every day;3=Every day</i>
IP3G24B	Provide information about an educational issue
IP3G24D	Maintain, organize and analyze data (e.g., with a spreadsheet or database)
IP3G24E	Prepare presentations
IP3G24J	Work with a learning management system (e.g., [Moodle, Blackboard, Edmodo])
IP3G24K	Manage staff (e.g., scheduling, professional development)
IP3G24L	Prepare the curriculum
IP3G24M	School financial management

Table 12.77: Item parameters, item fits, and WRMSD for the scale Principals' use of ICT for general school-related activities (P_ICTUSE)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IP3G24B	-0.391	-1.426	-0.164	1.590	1.216	1.259
IP3G24D	-0.522	-0.365	-0.601	0.966	1.005	1.019
IP3G24E	0.997	-1.641	-0.434	2.075	1.171	1.166
IP3G24J	-0.237	-0.363	-0.050	0.414	1.624	1.516
IP3G24K	-0.439	-0.044	-0.715	0.759	0.936	0.964
IP3G24L	0.615	-1.042	-0.194	1.236	1.133	1.174
IP3G24M	-0.023	-0.415	-0.262	0.678	1.265	1.209

Principals' use of ICT for school-related communication activities (P_ICTCOM)

Another scale was constructed from the statements in Question 24. School principals were asked to indicate how often they used ICT for various school-related communication activities. For each activity, they selected from the following response options: "Never," "Less than once a month," "At least once a month but not every week," "At least once a week but not every day," or "Every day." The responses to these items were used to derive a scale reflecting principals' use of ICT for school-related communication activities (P_ICTCOM). This scale includes items related to the frequency with which school principals communicate with teachers, education authorities, principals and senior staff in other schools, and parents using ICT.

Higher values on this scale indicate more frequent use of ICT for communication purposes in school administration and leadership. The scale is composed of 4 items, item wording is presented in Table 12.78. The IRT results are acceptable (Table 12.79). Both outfits and infits are relatively bigger compared with other principal scales but still close to 1, suggesting a general good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.78: Item wording for the scale Principals' use of ICT for school-related communication activities (P_ICTCOM)

Scale or item	Question or item wording
P_ICTCOM	How often do you use ICT for the following activities? <i>Recoded response categories for scaling: 0=Never;0=Less than once a month;1=At least once a month but not every week;2=At least once a week but not every day;3=Every day</i>
IP3G24F	Communicate with teachers in your school
IP3G24G	Communicate with education authorities
IP3G24H	Communicate with principals and senior staff in other schools
IP3G24I	Communicate with parents

Table 12.79: Item parameters, item fits, and WRMSD for the scale Principals' use of ICT for school-related communication activities (P_ICTCOM)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IP3G24F	-1.426	0.262	-1.124	0.862	1.285	1.182
IP3G24G	-0.123	-2.346	-0.140	2.486	1.313	1.421
IP3G24H	1.518	-1.920	-0.077	1.997	1.367	1.353
IP3G24I	0.031	-1.520	-0.442	1.962	1.399	1.351

The following five scales were constructed from the optional generative artificial intelligence (ChatGPT or similar) questions of the principal questionnaire.

Principals' views on negative impact of ChatGPT on students' learning (P_CIMNEG)

Two scales were constructed from Question 8 of the optional questionnaire. School principals were asked to indicate how likely they believed AI-based tools (e.g., ChatGPT or similar technologies) would lead to various negative consequences for student learning. For each potential impact, they selected from the following response options: "Very likely," "Somewhat likely," "Somewhat unlikely," or "Not likely at all." The responses to these items were used to derive a scale reflecting principals' views on the negative impact of AI-based tools (P_CIMNEG). This scale includes items related to concerns that AI tools may hinder students' ability to develop independent thinking, encourage academic dishonesty, provide misleading or biased information, create anxiety about AI influence, and foster student dependence on AI for learning.

Higher values on this scale indicate stronger concerns among school principals about the potential negative effects of AI-based tools on student learning. The scale is composed of 5 items, item wording is presented in Table 12.80. The IRT results are good (Table 12.81). Both outfits and infits are mostly close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.80: Item wording for the scale Principals' views on negative impact of ChatGPT on students' learning (P_CIMNEG)

Scale or item	Question or item wording
P_CIMNEG	How likely is it that the use of [ChatGPT or similar tools] will result in the outcomes listed below, with respect to students' learning at your school? <i>Recoded response categories for scaling: 3=Very likely;2=Somewhat likely;1=Somewhat unlikely;0=Not likely at all;</i>
IA3G08A	...make it difficult for students to develop a deep understanding of concepts
IA3G08C	... encourage students to submit work that is not their own
IA3G08E	...confuse students with false, misleading, or biased information
IA3G08J	... make students anxious about the influence of technology on the world
IA3G08K	... result in students becoming dependent on the tools rather than learning for themselves

Table 12.81: Item parameters, item fits, and WRMSD for the scale Principals' views on negative impact of ChatGPT on students' learning (P_CIMNEG)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IA3G08A	0.122	-1.706	0.078	1.628	0.949	0.949
IA3G08C	-0.479	-1.197	-0.406	1.603	1.036	1.002
IA3G08E	-0.333	-2.001	-0.157	2.158	0.968	0.959
IA3G08J	1.182	-2.367	0.330	2.037	1.268	1.286
IA3G08K	-0.492	-1.970	0.019	1.951	0.864	0.889

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' views on positive impact of ChatGPT on students' learning (P_CIMPOS)

Another scale was constructed from Question 8 of the optional questionnaire. School principals were asked to indicate how likely they believed AI-based tools (e.g., ChatGPT or similar technologies) would contribute to various positive outcomes for student learning. For each potential impact, they selected from the following response options: "Very likely," "Somewhat likely," "Somewhat unlikely," or "Not likely at all." The responses to these items were used to derive a scale reflecting principals' views on the positive impact of AI-based tools (P_CIMPOS). This scale includes items related to the perceived benefits of AI tools in enhancing students' interest in learning, improving logical sequencing in writing, supporting creativity by generating ideas, enhancing the quality of student work, and strengthening students' capacity to critically evaluate information.

Higher values on this scale indicate stronger positive perceptions among school principals regarding the potential benefits of AI-based tools for student learning. The scale is composed of 6 items, item wording is presented in Table 12.82. The IRT results are good (Table 12.83). Both outfits and infits are all close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.82: Item wording for the scale Principals' views on positive impact of ChatGPT on students' learning (P_CIMPOS)

Scale or item	Question or item wording
P_CIMPOS	How likely is it that the use of [ChatGPT or similar tools] will result in the outcomes listed below, with respect to students' learning at your school? <i>Recoded response categories for scaling: 3=Very likely;2=Somewhat likely;1=Somewhat unlikely;0=Not likely at all;</i>
IA3G08B	...help students develop a greater interest in learning
IA3G08F	... help students to develop logically sequenced research questions
IA3G08G	... help students to improve the quality of their written work
IA3G08H	... support students' creativity by generating ideas for them to consider
IA3G08L	... improve students' capacity to critically evaluate information
IA3G08M	... help students to refine research questions to obtain the most relevant information

Table 12.83: Item parameters, item fits, and WRMSD for the scale Principals' views on positive impact of ChatGPT on students' learning (P_CIMPOS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IA3G08B	-0.181	-3.149	-0.089	3.238	1.071	1.079
IA3G08F	-0.065	-3.032	-0.039	3.070	0.829	0.837
IA3G08G	-0.006	-2.586	-0.140	2.726	1.017	1.035
IA3G08H	-0.172	-2.773	-0.111	2.883	0.917	0.932
IA3G08L	0.531	-3.026	0.148	2.878	0.983	0.986
IA3G08M	-0.108	-2.583	-0.322	2.905	0.925	0.931

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' views on increased workload as a consequence of ChatGPT (P_CWORK)

Three scales were constructed from the statements in Question 9 of the optional questionnaire. School principals were asked to indicate how likely they believed AI-based tools (e.g., ChatGPT or similar technologies) would lead to increased workload for teachers and students. For each potential impact, they selected from the following response options: "Very likely," "Somewhat likely," "Somewhat unlikely," or "Not likely at all." The responses to these items were used to derive a scale reflecting principals' views on increased workload due to AI-based tools (P_CWORK). This scale includes items related to expectations that teachers will need specific professional learning on AI, students will need to learn how to use AI responsibly, teachers will need to monitor students' use of AI, and students will need to understand AI's limitations and potential biases.

Higher values on this scale indicate perceived increased likelihood among school principals about the additional responsibilities and workload AI may introduce for both teachers and students. The scale is composed of 7 items, item wording is presented in Table 12.84. The IRT results are acceptable (Table 12.85). Both outfits and infits are mostly close to 1 except for IA3G09Q, suggesting a general good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.84: Item wording for the scale Principals' views on increased workload as a consequence of ChatGPT (P_CWORK)

Scale or item	Question or item wording
P_CWORK	How likely is it that the use of [ChatGPT or similar tools] will have the following consequences on the work of teachers at your school? <i>Recoded response categories for scaling: 3=Very likely;2=Somewhat likely;1=Somewhat unlikely;0=Not likely at all;</i>
IA3G09A	Teachers will need specific professional learning support on the use of [ChatGPT or similar tools] to support teaching and learning
IA3G09F	Students will need to learn about how [ChatGPT or similar tools] are produced
IA3G09G	Teachers will need to monitor the degree to which their students depend on [ChatGPT or similar tools] to complete their classwork
IA3G09L	Students will need to learn about the potential benefits to society of the use of [ChatGPT or similar tools]
IA3G09M	Students will need to learn about the potential risks to society of the use of [ChatGPT or similar tools]
IA3G09Q	Teachers will need to find ways to assess student learning that prevent the students from using [ChatGPT or similar tools]
IA3G09R	Students will need to learn how to decide when to use and when not to use [ChatGPT or similar tools]

Table 12.85: Item parameters, item fits, and WRMSD for the scale Principals' views on increased workload as a consequence of ChatGPT (P_CWORK)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IA3G09A	-0.107	-1.531	-0.377	1.907	1.198	1.144
IA3G09F	0.615	-1.554	-0.599	2.153	1.197	1.178
IA3G09G	0.016	-1.421	-0.733	2.154	0.921	0.950
IA3G09L	-0.127	-1.715	-0.606	2.322	0.900	0.914
IA3G09M	-0.421	-1.018	-0.753	1.771	0.705	0.794
IA3G09Q	0.586	-1.742	-0.190	1.931	1.358	1.291
IA3G09R	-0.563	-1.508	-0.469	1.977	0.856	0.886

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' views on negative consequences of ChatGPT on teacher work (P_CONEG)

Another scale was constructed from Question 9 of the optional questionnaire. School principals were asked to indicate how likely they believed AI-based tools (e.g., ChatGPT or similar technologies) would lead to various negative consequences for teachers' work. For each potential impact, they selected from the following response options: "Very likely," "Somewhat likely," "Somewhat unlikely," or "Not likely at all." The responses to these items were used to derive a scale reflecting principals' views on the negative consequences of AI-based tools (P_CONEG). This scale includes items related to concerns that AI tools may undermine teachers' professional authority, reduce teachers' role in student learning, lead to a decline in student engagement with teachers, and diminish teachers' ability to assess students' original work.

Higher values on this scale indicate stronger concerns among school principals about the potential negative effects of AI-based tools on the teaching profession. The scale is composed of 5 items, item wording is presented in Table 12.86. The IRT results are acceptable (Table 12.87). Both outfits and infits are mostly close to 1 except for IA3G09B, suggesting a general good fit. Due to small sample



size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.86: Item wording for the scale *Principals' views on negative consequences of ChatGPT on teacher work* (P_CONEG)

Scale or item	Question or item wording
P_CONEG	How likely is it that the use of [ChatGPT or similar tools] will have the following consequences on the work of teachers at your school? <i>Recoded response categories for scaling: 3=Very likely;2=Somewhat likely;1=Somewhat unlikely;0=Not likely at all;</i>
IA3G09B	The use of [ChatGPT or similar tools] by students will make it difficult for teachers to judge whether or not the work submitted by students is their own
IA3G09E	The use of [ChatGPT or similar tools] will undermine the professional standing of teachers
IA3G09N	The use of [ChatGPT or similar tools] will result in teachers using material that includes inaccurate information
IA3G09O	The use of [ChatGPT or similar tools] will result in teachers using material that does not accurately represent the curriculum
IA3G09P	The use of [ChatGPT or similar tools] will result in teachers using material that does not represent good pedagogy in the [subject] they are teaching

Table 12.87: Item parameters, item fits, and WRMSD for the scale *Principals' views on negative consequences of ChatGPT on teacher work* (P_CONEG)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IA3G09B	-1.817	-2.344	-0.097	2.441	1.465	1.405
IA3G09E	1.127	-2.757	0.311	2.446	1.165	1.182
IA3G09N	-0.082	-2.780	0.069	2.711	0.830	0.837
IA3G09O	0.239	-2.646	0.276	2.369	0.771	0.781
IA3G09P	0.533	-2.736	0.063	2.672	0.853	0.860

Note: The IRT parameters were derived from the international pooled PCM model.

Principals' views on positive consequences of ChatGPT on teacher work (P_COPOS)

Another scale was constructed from Question 9 of the optional questionnaire, school principals were asked to indicate how likely they believed AI-based tools (e.g., ChatGPT or similar technologies) would contribute to various positive outcomes for teachers' work. For each potential impact, they selected from the following response options: "Very likely," "Somewhat likely," "Somewhat unlikely," or "Not likely at all." The responses to these items were used to derive a scale reflecting principals' views on the positive consequences of AI-based tools (P_COPOS). This scale includes items related to the perceived benefits of AI tools in making teachers' work more efficient, reducing administrative burdens, improving instructional quality, and enhancing teachers' ability to support student learning.

Higher values on this scale indicate stronger positive perceptions among school principals regarding the potential benefits of AI-based tools for teachers' professional responsibilities. The scale is composed of 4 items, item wording is presented in Table 12.88. The IRT results are good (Table 12.89). Both outfits and infits are mostly close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.88: Item wording for the scale *Principals' views on positive consequences of ChatGPT on teacher work* (P_COPOS)

Scale or item	Question or item wording
P_COPOS	How likely is it that the use of [ChatGPT or similar tools] will have the following consequences on the work of teachers at your school? <i>Recoded response categories for scaling: 3=Very likely;2=Somewhat likely;1=Somewhat unlikely;0=Not likely at all;</i>
IA3G09C	The use of [ChatGPT or similar tools] will make it easier for teachers to plan lessons
IA3G09D	The use of [ChatGPT or similar tools] will make it easier for teachers to create learning resources
IA3G09H	The use of [ChatGPT or similar tools] will make it easier for teachers to create individualized learning programs for their students
IA3G09J	Teachers will benefit from using [ChatGPT or similar tools] help them assess their students' work

Table 12.89: Item parameters, item fits, and WRMSD for the scale *Principals' views on positive consequences of ChatGPT on teacher work* (P_COPOS)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
IA3G09C	-0.146	-2.970	-0.019	2.989	0.849	0.865
IA3G09D	-0.515	-2.558	-0.479	3.037	0.741	0.776
IA3G09H	0.014	-2.889	-0.054	2.942	0.931	0.937
IA3G09J	0.647	-2.841	-0.180	3.021	1.072	1.070

Note: The IRT parameters were derived from the international pooled PCM model.

ICT Coordinator Scales

ICT coordinators' reports on availability of ICT resources at school (C_ICTRES)

In Question 4, ICT coordinators were asked to indicate the availability of various technology infrastructure resources in their schools. For each resource, they selected from the following response options: "Made available to teachers and students," "Made available to teachers only," "Made available to students only" or "Not made available." The responses to these items were used to derive a scale reflecting ICT coordinators' reports on the availability of ICT infrastructure and resources (C_ICTRES). This scale includes items related to internet access through the school network, school network storage space, instant messaging services for school-related use, video conferencing systems, and school intranet applications and workspaces.

Higher values on this scale indicate greater availability of ICT infrastructure and digital resources in schools. The scale is composed of 7 items, item wording is presented in Table 12.90. The IRT results are acceptable (Table 12.91). Both outfits and infits are mostly close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.90: Item wording for the scale ICT coordinators' reports on availability of ICT resources at school (C_ICTRES)

Scale or item	Question or item wording
C_ICTRES	Are the following technology infrastructure resources made available by your school to teachers and students? <i>Recoded response categories for scaling: 0=Not made available; 1=Made available only to students or Made available only to teachers; 2=Made available to teachers and students</i>
II3G04A	Access to the internet through the school network
II3G04C	Space on a school network to store files
II3G04E	Instant messaging service for school-related use
II3G04F	Video conferencing system (e.g. [Zoom, WebEx GoTo Meeting, Google Meet]) for school-related use
II3G04H	A school intranet with applications and workspaces
II3G04I	Remote access to the school network from home
II3G04J	Digital learning resources that can be accessed offline

Table 12.91: Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on availability of ICT resources at school (C_ICTRES)

Item	Delta	Tau.Cat1	Tau.Cat2	Outfit	Infit
II3G04A	-1.566	-0.537	0.537	0.803	0.840
II3G04C	-0.132	0.418	-0.418	1.036	1.037
II3G04E	0.277	1.003	-1.003	1.161	1.052
II3G04F	-1.797	-0.092	0.092	1.607	1.644
II3G04H	0.976	0.534	-0.534	1.099	1.059
II3G04I	1.589	0.067	-0.067	1.073	1.093
II3G04J	0.653	0.829	-0.829	1.333	1.217

ICT coordinators' reports on availability of ICT resources at school - tools (C_RESTOOL)

Two scales were constructed from Question 5. ICT coordinators were asked to indicate the availability of various software and digital learning tools in their schools. For each resource, they selected from the following response options: "Made available to teachers and students," "Made available to teachers only," "Made available to students only" or "Not made available." The responses to these items were used to derive a scale reflecting ICT coordinators' reports on the availability of ICT-based learning tools (C_RESTOOL). This scale includes items related to the availability of practice programs or apps where teachers determine content, single-user digital learning games, concept mapping software, data logging and monitoring tools, and simulations and modeling software.

Higher values on this scale indicate greater availability of digital learning tools to support teaching and student learning in schools. The scale is composed of 9 items, item wording is presented in Table 12.92. The IRT results are acceptable (Table 12.93). The majority of outfits and infits are close to 1 except for items II3G05I, II3G05J, II3G05K, suggesting an acceptable fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.92: Item wording for the scale ICT coordinators' reports on availability of ICT resources at school - tools (C_RESTOOL)

Scale or item	Question or item wording
C_RESTOOL	Are the following software resources made available by your school to teachers and students? <i>Recoded response categories for scaling: 0=Not made available;1=Made available only to students or Made available only to teachers; 2=Made available to teachers and students</i>
II3G05A	Practice programs or apps where teachers decide which questions are asked of students (e.g., [Quizlet, Kahoot, mathfessor], etc.)
II3G05B	Single-user digital learning games (e.g., [Mathletics])
II3G05G	Concept mapping software (e.g., [Inspiration, Webspiration], etc.)
II3G05H	Data logging and monitoring tools (e.g., [Logger Pro]) that capture real-world data digitally for analysis (e.g., speed, temperature, etc.)
II3G05I	Simulations and modelling software (e.g., [NetLogo])
II3G05J	Graphing or drawing software
II3G05K	e-portfolios (e.g., [VoiceThread])
II3G05N	Educational virtual reality and/or augmented reality apps (e.g., [The Body VR], [Google Earth VR], [Math Alive], etc.)
II3G05O	Adaptive learning systems (software that gathers and uses student data to deliver personalized resources and learning activities to address the individual needs of students)

Table 12.93: Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on availability of ICT resources at school - tools (C_RESTOOL)

Item	Delta	Tau.Cat1	Tau.Cat2	Outfit	Infit
II3G05A	-1.275	0.885	-0.885	1.208	1.073
II3G05B	0.047	1.870	-1.870	1.202	1.123
II3G05G	-0.313	1.534	-1.534	1.015	1.036
II3G05H	0.545	0.744	-0.744	0.862	0.934
II3G05I	0.936	1.502	-1.502	2.220	1.758
II3G05J	-1.432	1.161	-1.161	2.078	1.648
II3G05K	0.891	2.282	-2.282	2.988	1.722
II3G05N	-0.154	1.479	-1.479	1.434	1.273
II3G05O	0.756	1.397	-1.397	1.453	1.163

ICT coordinators' reports on availability of ICT resources at school - utility software (C_RESUTIL)

Another scale was constructed from Question 5. ICT coordinators were asked to indicate the availability of various general software resources in their schools. For each resource, they selected from the following response options: "Made available to teachers and students," "Made available to teachers only," "Made available to students only" or "Not made available." The responses to these items were used to derive a scale reflecting ICT coordinators' reports on the availability of general ICT software and resources (C_RESUTIL). This scale includes items related to the availability of word-processing software, presentation software, video and photo editing software, and digital textbooks.

Higher values on this scale indicate greater availability of essential ICT tools for teaching and learning in schools. The scale is composed of 4 items, item wording is presented in Table 12.94. The IRT results are not that good (Table 12.95). Half of the items have very small infits and outfits. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.94: Item wording for the scale ICT coordinators' reports on availability of ICT resources at school - utility software (C_RESUTIL)

Scale or item	Question or item wording
C_RESUTIL	Are the following software resources made available by your school to teachers and students? <i>Recoded response categories for scaling: 0=Not made available; 1=Made available only to students or Made available only to teachers; 2=Made available to teachers and students</i>
II3G05D	Word-processor software (e.g., [Microsoft Word, Apple Pages, Google Docs], etc.)
II3G05E	Presentation software (e.g., [Microsoft PowerPoint, Apple Keynote, Google Slides], etc.)
II3G05F	Video and photo software for capture and editing (e.g., [Windows Movie Maker, iMovie, Adobe Photoshop], etc.)
II3G05M	Digital textbooks

Table 12.95: Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on availability of ICT resources at school - utility software (C_RESUTIL)

Item	Delta	Tau.Cat1	Tau.Cat2	Outfit	Infit
II3G05D	-0.726	0.870	-0.870	0.476	0.692
II3G05E	-0.844	0.829	-0.829	0.377	0.676
II3G05F	0.655	1.000	-1.000	1.231	1.109
II3G05M	0.915	0.591	-0.591	1.003	1.021

Note: The IRT parameters were derived from the international pooled PCM model.

ICT coordinators' reports on computer resource hindrances to the use of ICT in teaching and learning (C_HINRES)

In Question 14, ICT coordinators were asked to indicate the extent to which various technical factors hinder the use of ICT in teaching and learning. For each factor, they selected from the following response options: "A lot," "To some extent," "Very little," or "Not at all." The responses to these items were used to derive a scale reflecting ICT coordinators' reports on technical hindrances to ICT use (C_HINRES). This scale includes items related to insufficient numbers of computers with internet access, inadequate internet bandwidth or speed, lack of sufficiently powerful computers, problems with ICT equipment maintenance, and insufficient computer software availability.

Higher values on this scale indicate greater perceived barriers to ICT integration due to technical resource constraints. The scale is composed of 7 items, item wording is presented in Table 12.96. The IRT results are good (Table 12.97). Both outfits and infits are mostly close to 1, suggesting a good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.

Table 12.96: Item wording for the scale ICT coordinators' reports on computer resource hindrances to the use of ICT in teaching and learning (C_HINRES)

Scale or item	Question or item wording
C_HINRES	To what extent is the use of ICT in teaching and learning at your school hindered by each of the following obstacles? <i>Recoded response categories for scaling: 0=Not at all;1=Very little;2=To some extent;3=A lot</i>
II3G14A	Not enough computers with an internet connection
II3G14B	Insufficient internet bandwidth or speed
II3G14D	Not enough sufficiently powerful computers
II3G14E	Problems in maintaining ICT equipment
II3G14F	Not enough computer software
II3G14H	Inefficient technical ICT support (e.g., provision of new resources takes too long, the time to fix technical problems is too long)
II3G14I	Not enough screens for displaying digital content (e.g., TVs, projectors, large display monitors, smartboards, etc.)

Table 12.97: Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on computer resource hindrances to the use of ICT in teaching and learning (C_HINRES)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
II3G14A	0.415	-0.466	-0.280	0.746	0.971	0.991
II3G14B	0.135	-0.770	-0.111	0.881	1.367	1.262
II3G14D	-0.175	-0.830	-0.496	1.326	0.994	1.028
II3G14E	-0.250	-1.320	-0.212	1.533	0.990	1.013
II3G14F	-0.136	-1.415	-0.296	1.712	1.065	1.097
II3G14H	-0.127	-1.426	0.300	1.126	1.182	1.181
II3G14I	0.137	-0.938	0.017	0.922	1.053	1.087

ICT coordinators' reports on pedagogical resource hindrances to the use of ICT in teaching and learning (C_HINPED)

In Question 19, ICT coordinators were asked to indicate the extent to which various pedagogical factors hinder the use of ICT in teaching and learning. For each factor, they selected from the following response options: "A lot," "To some extent," "Very little," or "Not at all." The responses to these items were used to derive a scale reflecting ICT coordinators' reports on pedagogical resourcing hindrances to ICT use (C_HINPED). This scale includes items related to insufficient ICT skills among teachers, lack of time for lesson preparation, inadequate professional learning resources, ineffective online learning support, and restrictive school ICT policies.

Higher values on this scale indicate greater perceived barriers to ICT integration due to pedagogical resourcing constraints. The scale is composed of 9 items, item wording is presented in Table 12.98. The IRT results are acceptable (Table 12.99). Both outfits and infits are mostly close to 1 except for item II3G19J, suggesting a general good fit. Due to small sample size, there is no by country analysis for principal scales, therefore no measurement invariance results are available across countries.



Table 12.98: Item wording for the scale ICT coordinators' reports on pedagogical resource hindrances to the use of ICT in teaching and learning (C_HINPED)

Scale or item	Question or item wording
C_HINPED	To what extent is the use of ICT in teaching and learning at your school hindered by each of the following obstacles? <i>Recoded response categories for scaling: 0=Not at all;1=Very little;2=To some extent;3=A lot</i>
II3G19A	Insufficient ICT skills among teachers
II3G19B	Insufficient time for teachers to prepare lessons
II3G19C	Insufficient effective professional learning resources for teachers
II3G19D	Lack of an effective online learning support platform
II3G19E	Insufficient incentives for teachers to integrate ICT use in their teaching
II3G19F	Restricted access to useful Internet resources
II3G19G	Insufficient pedagogical support for the use of ICT
II3G19I	Lack of a school-wide vision for using ICT to support teaching and learning
II3G19J	The school's ICT use policy prevents the use of valuable ICT resources that would support teaching and learning

Table 12.99: Item parameters, item fits, and WRMSD for the scale ICT coordinators' reports on pedagogical resource hindrances to the use of ICT in teaching and learning (C_HINPED)

Item	Delta	Tau.Cat1	Tau.Cat2	Tau.Cat3	Outfit	Infit
II3G19A	-1.336	-2.647	-0.031	2.678	1.169	1.160
II3G19B	-1.081	-1.457	-0.275	1.732	1.386	1.277
II3G19C	-0.322	-1.713	-0.313	2.027	0.895	0.902
II3G19D	-0.037	-1.789	0.134	1.655	1.142	1.159
II3G19E	-0.660	-1.479	-0.256	1.735	1.005	1.003
II3G19F	0.737	-1.467	0.111	1.357	1.156	1.145
II3G19G	0.258	-1.832	-0.106	1.938	0.942	0.961
II3G19I	0.699	-1.719	-0.150	1.869	1.274	1.291
II3G19J	1.742	-1.449	0.115	1.334	1.476	1.397

References

- Buchholz, J., & Hartig, J. (2019). Comparing attitudes across groups: An irt-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement*, 43(3), 241–250. <https://doi.org/10.1177/0146621617748323>
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Fraillon, J., & Rožman, M. (2024). *IEA International Computer and Information Literacy Study 2023 assessment framework*. Springer. <https://doi.org/10.1007/978-3-031-61194-0>
- Ganzeboom, H. B., de Graaf, P. M., & Treiman, D. J. (1992). A standard international socioeconomic index of occupational status. *Social Science Research*, 21, 1–56.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer.
- International Labour Organization. (2007). *International standard classification of occupations: ISCO-2008*. International Labour Office.
- Jolliffe, I. T. (2002). *Principal component analysis* (2nd). Springer.



- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–122). Springer.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd). McGraw-Hill.
- Robitzsch, A., Kiefer, T., & Wu, M. (2024). TAM: Test analysis modules. <https://CRAN.R-project.org/package=TAM>
- UNESCO Institute for Statistics. (2012). *International Standard Classification of Education: ISCED 2011*. UNESCO-UIS. <https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf>
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. <https://doi.org/10.1007/bf02294627>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.
- Yamamoto, K., Khorramdel, L., Von Davier, M., et al. (2013). *Scaling PIAAC cognitive data* (tech. rep.). OECD Publishing Paris.

Chapter 13:

Reporting the ICILS 2023 results

Andrés Strello, Mojca Rožman, and Julian Fraillon

Introduction

This chapter describes the procedures for obtaining the population estimates presented in the ICILS 2023 international report. The chapter is divided into three sections. In the first section, we explain how the sampling design affects the estimation of population parameters. We first consider the use of weights to attain unbiased estimations of these parameters and then discuss how to quantify the uncertainty in the estimation arising from the complex sampling design with which the data was generated. In the second section, we describe the imputation variance that quantifies the uncertainty arising from the test design. This uncertainty arises from using balanced incomplete block designs for the computer and information literacy (CIL) and computational thinking (CT) tests, where each student completes only a subset of items within the total item pool. In the subsequent section, we describe how the relationship between CIL and CT scores was calculated. Finally, we describe the procedures for conducting significance tests for differences between averages, country, and subgroup means or percentages, and between results from ICILS 2023, 2018, and 2013 for the countries that participated in more than one cycle.

13.1 Estimation of sampling variance

Sampling, weights, and non-response model

In order to select schools, teachers, and students, ICILS 2023 used a stratified two-stage probability sampling design. In order to facilitate the estimation of population parameters accounting for the probabilistic nature of the sampling design and the non-response model, the ICILS 2023 databases provide a set of total sampling weights. These variables are the product of the varying probabilities of inclusion across participating units and the different non-response adjustment factors. For more details on the non-response adjustments and the construction of the sampling weights, see [Chapter 6](#) and [Chapter 7](#), respectively. Each unit in the ICILS 2023 databases is accompanied by a total sampling weight variable: TOTWGTC for the school-level, TOTWGTS for students, and TOTWGTT for teachers. Weighted estimation using these variables is necessary to obtain unbiased population parameter estimates.

Jackknife repeated replication

The complex sampling design requires the estimation of variance to account for the following three features. The first is explicit stratification, which follows that an estimated parameter's sampling variance is the sum of the stratum-level sampling variances. The second is that the total sampling weights may vary across units. Finally, the two-stage process of the sampling design implies that student and teacher characteristics or traits are not independent and identically distributed because they are clustered within schools. To account for these features in the estimation, ICILS 2023 (as this was the case in previous ICILS cycles and in other IEA studies) uses the *jackknife repeated replication* (JRR) variance estimator. Replication (re-sampling) techniques are useful tools to approximate the sampling variance of population estimates (Gonzalez & Foy, 2000; Quenouille, 1949; Tukey, 1958; Wolter, 1985). ICILS 2023 uses the JRR variance estimator when reporting standard errors for all estimated population parameters, such as means, percentages, correlation coefficients, and regression parameters.

During sample selection, schools are sorted within each explicit stratum by a measure of their size and, where appropriate and feasible, other observable characteristics so that adjacent schools are often more similar to each other. In the JRR estimation procedure, pairs of adjacent primary sampling units



(PSUs)—typically sampled schools—are allocated to pseudo strata. These pseudo-strata are referred to as jackknife zones or variance strata. In the publicly available datasets, these are captured by the variable JKZONE, with an additional character identifying the relevant population.²⁸ Some specific rules apply when there is an odd number of PSUs or if the number of PSUs exceeds 150. Please refer to [Chapter 6](#) for more details about this. Each country participating in ICILS 2023 has up to 75 jackknife zones available for each population (see [Table 13.1](#)).

Within each jackknife zone, the PSUs were assigned to two different groups. This assignment is documented in the jackknife replicate code variable JKREP, where the groups are assigned either the value 1 or 0. Based on the jackknife zone and jackknife replicate code, a total of R different replication weights were created (RWGT). Each replication weight is derived by manipulating the total sampling weights of all sampling units in a respective jackknife zone. That is, to create the $r = 1, \dots, R$ replicate weight, the total sampling weight of all units is either doubled (in case PSUs within jackknife zone r have a jackknife replicate code of one) or set to zero (in case of a jackknife replicate code of zero). The total sampling weights of all units falling outside the jackknife zone r are kept unmodified. A total of R replicate weights are computed by iterating across all R jackknife zones. If fewer than 75 zones were in a dataset, the remaining replicate weights were set to the original sampling weight. Therefore, they do not contribute to the sampling variable estimate.

[Table 13.2](#) shows an example of the variables included in a dataset that are needed to calculate the sampling variance.²⁹ There are a total of five jackknife zones in this example. In the first column, we show the school ID; in the second and third columns, the jackknife zone and jackknife zone replicate code, respectively; in the fourth, the total sampling weight (TOTWGTC). The replicate weights 1-5 follow in the remaining of the columns. The respondents within schools with ID 1 or ID 2 are part of the jackknife zone 1, with a sampling weight of 48.96 and 25.77, respectively. The first school has a replicate code of zero, and the second has a replicate code of one. Therefore, the weight of the first school will be zeroed, and the weight of the second school will be doubled when generating the first replicate weight. The total weight remains unchanged for all schools outside the first jackknife zone. The same procedure is applied for the schools in zone 2 for the second replicate variable. The value for replicate weight 2 for the school with a code zero will be zeroed, and the value for the other school will be doubled, whereas the rest of the schools will retain their original total weight. This procedure continues until the last jackknife zone.

The last row in [Table 13.2](#) presents the weighted means of the variable C_ICTSTD (sum of ICT devices available for student use) using different sets of weights: total sampling weight (which would correspond to the reported mean in the ICILS 2023 international report), and for each of the five replicate weights (i.e., replicated means). The sampling variance of the estimated mean is, therefore, approximated by measuring the variation between the reported mean and the replicated means.

28 JKZONES represent sampling zones for students, JKZONEC for schools, and JKZONET for teachers. The same logic applies to JKREP and RWGT mentioned below.

29 Please note that [Table 13.2](#) uses fictional school IDs and jackknife zones, so these numbers do not correspond to the available datasets.



Table 13.1: Number of jackknife zones in national samples

Country	School data	Student data	Teacher data
Austria	75	75	68
Azerbaijan	75	75	74
Belgium (Flemish)	69	69	64
Bosnia and Herzegovina	46	57	19
Chinese Taipei	75	75	75
Croatia	75	74	75
Cyprus	49	75	74
Czech Republic	75	75	75
Denmark	69	71	63
Finland	75	75	75
France	68	75	61
Germany	75	75	75
Greece	75	75	75
Hungary	75	75	72
Italy	75	75	75
Kazakhstan	75	75	75
Korea, Republic of	75	75	75
Kosovo	72	75	57
Latvia	73	73	69
Luxembourg	18	75	11
Malta	21	75	16
Netherlands	20	23	16
Norway (Grade 9)	74	75	70
Oman	75	75	75
Portugal	75	75	75
Romania	64	70	67
Serbia	75	75	75
Slovak Republic	75	75	75
Slovenia	75	75	75
Spain	75	75	75
Sweden	75	75	73
United States	63	61	50
Uruguay	73	73	65
Benchmarking participant			
North Rhine-Westphalia, Germany	57	57	54



Table 13.2: Example of computation of jackknife replicate weights

School ID	Jackknife zone	Jackknife replicate code	C_ICTSTD	Total sampling weight	Replicate weight 1	Replicate weight 2	Replicate weight 3	Replicate weight 4	Replicate weight 5
1	1	0	111	48.96	00.00	48.96	48.96	48.96	48.96
2	1	1	51	25.77	51.54	25.77	25.77	25.77	25.77
3	2	0	25	40.08	40.08	00.00	40.08	40.08	40.08
4	2	1	47	40.08	40.08	80.15	40.08	40.08	40.08
5	3	0	198	05.17	05.17	05.17	00.00	05.17	05.17
6	3	1	120	05.04	05.04	05.04	10.08	05.04	05.04
7	4	0	48	16.23	16.23	16.23	16.23	00.00	16.23
8	4	1	54	20.42	20.42	20.42	20.42	40.83	20.42
9	5	0	95	06.18	06.18	06.18	06.18	06.18	00.00
10	5	1	120	05.55	05.55	05.55	05.55	05.55	11.09
Weighted mean of C_ICTSTD:				67.45	54.01	71.58	65.52	67.64	68.02

Note: The table is based on the school dataset. C_ICTSTD: sum of ICT devices available for student use. Because of rounding some results may appear inconsistent.

The sampling variance of an estimated parameter $\hat{\theta}$ is approximated using the formula:

$$SV_{\hat{\theta}} = \sum_{r=1}^R (\hat{\theta}_r - \hat{\theta})^2, \quad (13.1)$$

where $\hat{\theta}$ corresponds to the reported estimate of θ (i.e., estimated using total sampling weights) and $\hat{\theta}_r$ corresponds to the estimate of the parameter using the replicate weight r . As above, R corresponds to the total number of jackknife zones included in the dataset, which is 75 in ICILS 2023.

The standard error $SE_{\hat{\theta}}$ corresponds to the square root of the sampling variance $SV_{\hat{\theta}}$:

$$SE_{\hat{\theta}} = \sqrt{SV_{\hat{\theta}}}. \quad (13.2)$$

In the example presented in Table 13.2, that includes the weighted means of the variable C_ICTSTD, the standard error would be calculated as follows. The mean in the column of the total weight (67.45) is representing $\hat{\theta}$, while $\hat{\theta}_r$ corresponds to each of the means in the columns of replicate weights. Following this, the sampling variance would correspond to 201.62, which leads to a standard error of 14.20.

Sampling variance estimation using JRR can be used when estimating parameters, such as means, percentages, standard deviations, correlations, and differences in means, for data that do not include student achievement scores. In ICILS 2023, these relate to data collected using the questionnaires. When working with estimates of CIL or CT scores, imputation variance must also be included. In the next section, we explain how the imputation variance is incorporated into the standard error of estimated parameters involving CIL and CT achievement scores.

13.2 Estimation of imputation variance for CIL and CT scores

ICILS 2023 includes two tests of student achievement—one for CIL and one for CT. Each test instrument contains a pool of items delivered using a balanced incomplete block design (see Chapter 2 for details of the test development and test design) in which students completed a subset of CIL and (where relevant) CT test items. Having students respond to a subset of items introduces additional



uncertainty to the measures of student achievement. In order to accommodate this uncertainty, for ICILS 2023, each student's score (CIL and—where applicable CT) was represented by five random draws from an individual's posterior distribution—plausible values—which represent the uncertainty in the measurement of student achievement (for more details on plausible values see [Chapter 11](#)). When reporting estimates of parameters characterizing the distribution of CIL and CT scores in a population, the variation among these values must be considered (von Davier et al., 2009). Population estimates and their standard errors for ICILS 2023 CIL and CT scores should always be estimated using all five plausible values. This procedure follows the principles developed by Rubin (1987), where the estimation is performed once per each plausible value and the results are aggregated (Mislevy et al., 1992). For more information about plausible values and their use in secondary analyses, see, for example, Jewsbury et al. (2024).

To estimate a parameter θ that can be used to characterize the distribution of CIL or CT scores in a (sub)population, the estimate of interest $\hat{\theta}$ can be computed as the average across the estimates using each plausible value:

$$\hat{\theta} = \frac{1}{P} \sum_{p=1}^P \hat{\theta}^p, \quad (13.3)$$

where $\hat{\theta}^p$ corresponds to the estimate of θ based on plausible value $p = 1, \dots, P$ and $P = 5$, corresponding to the five plausible values of CIL and CT that are drawn in ICILS.

The total variance of $\hat{\theta}$ is a composite of the sampling and the imputation variance. The sampling variance $SV_{\hat{\theta}}$ is calculated as the average of the sampling variances across the plausible values:

$$SV_{\hat{\theta}} = \frac{1}{P} \sum_{p=1}^P SV_{\hat{\theta}}^p, \quad (13.4)$$

where $SV_{\hat{\theta}}^p$ corresponds to the sampling variance of $\hat{\theta}$ based on the p^{th} plausible value, similar to the one documented in [13.1](#).

The imputation variance $IV_{\hat{\theta}}$ consists of the variation across $\hat{\theta}^p$ relative to its mean; that is,

$$IV_{\hat{\theta}} = \frac{1}{P-1} \sum_{p=1}^P (\hat{\theta}^p - \hat{\theta})^2. \quad (13.5)$$

The total variance $TV_{\hat{\theta}}$ is then the sum of the sampling and the imputation variances:

$$TV_{\hat{\theta}} = SV_{\hat{\theta}} + \left(1 + \frac{1}{P}\right) IV_{\hat{\theta}}. \quad (13.6)$$

Finally, the reported standard error of $\hat{\theta}$ is given by:

$$SE_{\hat{\theta}} = \sqrt{TV_{\hat{\theta}}}. \quad (13.7)$$

The same approach was followed whenever CIL or CT scores were used in the analyses. For example, to calculate means, mean differences, correlation, or regression coefficients with the corresponding standard errors, among others. The sampling and imputation components of the variance for national CIL and CT scores, together with the national sample size, are presented in [Table 13.3](#) and [Table 13.4](#).



Table 13.3: CIL mean variance components and reported standard error per country

Country	CIL mean score	N	Sampling variance	Imputation variance	Total variance	Standard error
Austria	505.58	3,448	5.71	0.55	6.37	2.52
Azerbaijan	318.65	3,634	23.25	2.27	25.97	5.10
Belgium (Flemish)	510.61	3,365	19.14	0.42	19.65	4.43
Bosnia and Herzegovina	440.28	1,877	13.56	0.65	14.34	3.79
Chinese Taipei	515.27	5,112	9.04	0.09	9.15	3.02
Croatia	486.60	2,911	14.10	0.67	14.90	3.86
Cyprus	460.42	3,182	6.51	0.13	6.66	2.58
Czech Republic	525.39	8,169	4.14	0.21	4.39	2.10
Denmark	517.86	3,038	6.68	0.38	7.13	2.67
Finland	506.69	4,249	12.52	0.12	12.66	3.56
France	497.64	3,694	7.22	0.07	7.31	2.70
Germany	501.84	5,065	11.53	0.40	12.00	3.46
Greece	459.91	3,576	11.09	0.10	11.21	3.35
Hungary	504.94	3,491	14.51	0.17	14.71	3.84
Italy	490.54	3,376	6.37	0.18	6.59	2.57
Kazakhstan	406.99	4,852	9.31	0.28	9.64	3.10
Korea, Republic of	540.40	3,723	6.27	0.06	6.35	2.52
Kosovo	355.55	3,345	16.34	0.28	16.67	4.08
Latvia	508.81	2,705	12.79	0.39	13.26	3.64
Luxembourg	493.67	4,703	3.88	0.04	3.93	1.98
Malta	475.02	3,115	5.68	0.59	6.39	2.53
Netherlands	459.58	1,288	100.95	0.53	101.59	10.08
Norway (Grade 9)	502.07	4,436	7.85	0.24	8.14	2.85
Oman	379.12	6,437	8.80	0.07	8.88	2.98
Portugal	509.71	3,650	8.71	0.38	9.17	3.03
Romania	417.59	3,270	27.88	0.19	28.10	5.30
Serbia	443.00	3,125	13.58	0.14	13.74	3.71
Slovak Republic	498.69	3,034	6.79	0.53	7.43	2.73
Slovenia	483.48	3,318	4.85	0.31	5.22	2.29
Spain	495.20	11,799	3.55	0.14	3.72	1.93
Sweden	504.24	3,401	8.69	0.47	9.26	3.04
United States	481.64	2,352	42.60	0.29	42.94	6.55
Uruguay	447.13	2,933	12.93	0.20	13.17	3.63
Benchmarking participant						
North Rhine-Westphalia, Germany	484.71	2,726	15.88	0.89	16.95	4.12



Table 13.4: CT variance components and reported standard error by country

Country	CT mean score	N	Sampling variance	Imputation variance	Total variance	Standard error
Austria	476.32	3,448	14.00	1.03	15.23	3.90
Belgium (Flemish)	509.12	3,365	38.42	1.18	39.84	6.31
Chinese Taipei	548.40	5,112	14.88	0.18	15.09	3.88
Croatia	428.71	2,911	17.41	1.30	18.97	4.36
Czech Republic	526.69	8,169	8.03	0.16	8.22	2.87
Denmark	503.97	3,038	12.06	0.20	12.30	3.51
Finland	501.50	4,249	25.98	0.65	26.75	5.17
France	499.32	3,694	14.44	0.42	14.95	3.87
Germany	479.03	5,065	14.43	0.30	14.78	3.85
Italy	482.02	3,376	8.65	0.23	8.93	2.99
Korea, Republic of	536.75	3,723	10.29	0.36	10.73	3.28
Latvia	495.20	2,705	26.28	0.70	27.12	5.21
Luxembourg	475.52	4,703	5.89	0.25	6.19	2.49
Malta	437.94	3,115	9.60	0.10	9.71	3.12
Netherlands	440.33	1,288	164.03	0.48	164.61	12.83
Norway (Grade 9)	484.72	4,436	13.38	0.43	13.89	3.73
Portugal	483.88	3,650	15.80	0.33	16.19	4.02
Serbia	421.76	3,125	25.40	0.61	26.14	5.11
Slovak Republic	498.15	3,034	12.17	1.29	13.71	3.70
Slovenia	448.25	3,318	9.43	0.66	10.22	3.20
Sweden	486.43	3,401	21.28	1.19	22.72	4.77
United States	461.45	2,352	42.72	6.09	50.03	7.07
Uruguay	421.10	2,933	17.86	0.20	18.10	4.25
Benchmarking participant						
North Rhine-Westphalia, Germany	460.69	2,726	16.57	0.19	16.81	4.10

13.3 Reporting relationships between CIL and CT scores

There are two examples in the ICILS 2023 international report that require some additional clarification on how to be computed. The first is the correlation between CIL and CT, i.e., where the two sets of plausible values are correlated. The other relates to the estimation of student CT mean achievement at each of the CIL proficiency levels³⁰ (both included in Table 5.8 in the ICILS 2023 international report).

Reporting correlation between CIL and CT scores

To understand how plausible values from different domains can be used in the same analysis, it is useful to know how they are generated. For countries using both domains, plausible values for CIL and CT are drawn simultaneously. We start by drawing the first CIL and CT plausible values, then the second for both domains and so on (for more details, see [Chapter 11](#)). Therefore, the correlation between the first plausible values of two domains provides an unbiased estimate of the relationship between the domains (Jewsbury et al., 2024). To obtain the estimates of the correlation coefficient between CIL and CT that were reported in Table 5.8 in Chapter 5 of the ICILS 2023 international report, the correlation coefficient between CIL and CT for each country was estimated P times (which is in our case five times), each time using CIL and CT plausible value p . The reported estimate and its standard error are then estimated using Equations (13.3) and (13.6).

³⁰ ICILS established a described scale, where certain cut-off points are used to define CIL and CT proficiency levels (for more details on proficiency levels, see [Chapter 11](#)).



Reporting student CT mean achievement by CIL proficiency levels

To estimate the mean CT achievement by CIL proficiency levels (see Table 5.8 in the international report), each student must be categorized into a proficiency level according to their CIL score on each plausible value. Therefore, it may be that a student is categorized as reaching proficiency level *A* when using one plausible value while categorized as reaching proficiency level *B* when considering a different plausible value. This student will contribute to the estimate of mean CT achievement among those students reaching proficiency level *A* of CIL when using one plausible value. In contrast, they will not contribute to the same estimate when considering the next plausible value. In summary, the mean CT of students reaching proficiency level *A* in CIL is estimated at *P* times. Only students reaching proficiency level *A* when using the CIL plausible value *p* will contribute with their CT plausible value *p* to the estimate of the mean CT score for this CIL proficiency level.

In a second step, the reported estimate of mean CT achievement among students reaching CIL proficiency level *A* and its standard error are estimated as described in the previous sections (see Equations (13.3) and (13.6)).

13.4 Reporting of differences

In this section, we explain how we calculated the standard error of group differences in the ICILS 2023 international report. Four types of group differences were used: differences between countries (e.g., difference in CIL between two countries), differences between subgroups of students within countries (e.g., gender difference in CIL within countries), differences between countries and an international average (e.g., difference between a country CIL and the ICILS 2023 CIL average), and differences between cycles within a country (e.g., the difference in CIL in a country between years 2023 and 2018).

The estimate of these differences $\hat{\theta}_{(a-b)}$ can be generalized as:

$$\hat{\theta}_{(a-b)} = \hat{\theta}_a - \hat{\theta}_b, \quad (13.8)$$

that is, a simple difference between $\hat{\theta}_a$ and $\hat{\theta}_b$. The t-statistic associated to this estimated difference is then given by:

$$t_{(a-b)} = \frac{\hat{\theta}_{(a-b)}}{SE_{\hat{\theta}_{(a-b)}}}, \quad (13.9)$$

where $SE_{\hat{\theta}_{(a-b)}}$ corresponds to the standard error of the estimated difference. Recall, that as a general rule,

$$Var_{\hat{\theta}_{(a-b)}} = Var_{\hat{\theta}_a} + Var_{\hat{\theta}_b} - 2 * Cov(\hat{\theta}_a, \hat{\theta}_b), \quad (13.10)$$

therefore, the estimation of $SE_{\hat{\theta}_{(a-b)}}$ depends on whether or not, by design, the sampling covariance of both estimates exists. If the data from which both estimates are derived are generated with the same random process, then we say there is a sampling covariance between the estimates. This would be the case, for example, of boys and girls: in most countries, boys and girls attend the same schools, therefore, the subsample of boys and subsample of girls are based on the same sample of schools, i.e., both subsamples are not independent. We typically refer to this by saying that the samples are dependent. However, if each estimate is derived using data generated with an independent random process, then, by design, the sampling covariance between both estimates is zero. As example, the sample of different countries are drawn separately and do not depend on each other. We typically refer to this by saying that the samples are independent. Below, we discuss different scenarios for estimating $SE_{\hat{\theta}_{(a-b)}}$, emphasizing whether or not the estimate of sampling covariance should be included.

In addition, it is worth noting that in the ICILS 2023 international report, we highlight differences that are statistically different from zero with a confidence level of 95 percent. To do so, we compare $t_{(a-b)}$ with the critical values of the t-distribution. Since we have large samples, this distribution is equivalent to the standard normal distribution.



Differences in population estimates between countries

Appendix C of the international report includes countries' differences in mean CIL and CT scores. Since for each country (or explicit stratum within a country), an independent random process selects students to participate in ICILS, it follows that there is no sampling covariance between the estimates of two different countries. Therefore, the variance of the difference between the estimates of country c and d is defined as the sum of the variance of the estimate of each country. The standard error of the estimated difference in this context can be expressed as:

$$SE_{(c-d)} = \sqrt{SE_c^2 + SE_d^2}. \quad (13.11)$$

Differences between subgroups within countries

In most cases, when we estimated differences between subgroups of the target population within a country, we had to consider that the sampling covariance between those estimates is larger than zero. That is, the samples contributing to each of the estimates are dependent. An exception to this is when the subgroups are in different explicit strata, in which case the sampling covariance is zero (for the description of the national samples and the national explicit strata, please see [Appendix B](#)). In this latter case, the samples are independent, and the methodology used to estimate the standard error of the difference follows the methodology used when we estimate differences between two countries (see below).

Following, we describe how we estimated the standard error of an estimated difference when sampling covariance between the two estimates exists. As explained in [Section 13.1](#) and [Section 13.2](#), the computation of the standard error depends on whether the estimate involves plausible values (i.e., CIL or CT) or not.

Standard error of differences between estimates based on dependent samples not involving plausible values

The standard error of differences between estimates of dependent samples must consider the sampling covariance of the two estimates. Considering [\(13.2\)](#), and assuming that the two groups are named as a and b , the standard error of the difference estimate $SE_{(a-b)}$ is computed as:

$$SE_{(a-b)} = \sqrt{\sum_{r=1}^R ((a_r - b_r) - (a_0 - b_0))^2}, \quad (13.12)$$

where a_r and b_r are the estimates of subgroups a and b respectively per each replication weight r , while a_0 and b_0 are the estimates of subgroups a and b using the total sampling weight. In other words, the sampling variance consists of the variation of the difference between both groups among jackknife replications.

Standard error of differences between estimates based on dependent samples involving plausible values

As explained in [Section 13.2](#), any parameter estimated involving the variables CIL or CT needs to consider the uncertainty arising from the implemented assessment design. That is, any estimate should consider the variation in the estimate arising from using the different plausible values. In the context of CIL and CT, the sampling variance of the difference, $SV_{(a-b)}$, can be formulated as:

$$SV_{(a-b)} = \frac{1}{P} \sum_{p=1}^P \left(\sum_{r=1}^R ((a_{r,p} - b_{r,p}) - (a_{0,p} - b_{0,p}))^2 \right), \quad (13.13)$$

where P corresponds to the total number of plausible values (five in ICILS 2023); $a_{r,p}$ and $b_{r,p}$ corresponds to the subgroups a and b estimates for plausible value p and replication weight r , respectively; and, $a_{0,p}$ and $b_{0,p}$ corresponds to the subgroups estimates for plausible value p and using the total sampling weights.



The imputation variance of the difference $IV_{(a-b)}$ can be formulated as:

$$IV_{(a-b)} = \frac{\sum_{p=1}^P ((a_{0,p} - b_{0,p}) - (\bar{a}_{0,P} - \bar{b}_{0,P}))^2}{P - 1}, \quad (13.14)$$

where $(\bar{a}_{0,P} - \bar{b}_{0,P})$ is defined as:

$$(\bar{a}_{0,P} - \bar{b}_{0,P}) = \frac{\sum_{p=1}^P (a_{0,p} - b_{0,p})}{P}. \quad (13.15)$$

Finally, the total variance of the difference is composed by the sum of the sampling variance and the imputation variance, similar to (13.6). Thus, the standard error of the difference can be formulated as:

$$SE_{(a-b)} = \sqrt{SV_{(a-b)} + \left(1 + \frac{1}{P}\right) IV_{(a-b)}} \quad (13.16)$$

Differences between a country and the ICILS 2023 average

In the ICILS 2023 international report, most of the analyses compare a country estimate (e.g., mean, estimated difference, percentage, among others) to the average of the same estimate across ICILS 2023 countries. The ICILS 2023 averages α were calculated as follows:

$$\alpha = \frac{\sum_{c=1}^C \hat{\theta}_c}{C}, \quad (13.17)$$

where $\hat{\theta}_c$ corresponds to the estimate of country $c = 1, \dots, C$ and C corresponds to the total number of countries involved in the ICILS 2023 average estimate. All non-benchmarking participants that meet sampling participation requirements except Romania were considered for the ICILS 2023 average because they collected their data later than the other participants. This means that in most cases where an international average was estimated using CIL, C corresponded to 30 countries, while in analysis involving CT, C corresponded to 21 countries. However, this number could have varied, as in some specific tables, some countries were excluded for other reasons (e.g., a low number of valid cases per subgroup or the question not being applied in that country).

The standard error of an estimated (international) ICILS 2023 parameter was computed as the square root of the sum overall all country-specific sampling variances, that is:

$$SE_{\alpha} = \sqrt{\frac{\sum_{c=1}^C SE_{\theta_c}^2}{C^2}}, \quad (13.18)$$

where SE_{θ_c} corresponds to the standard error of the country c estimate.

Differences with countries not included in the average

Several tables presented in the ICILS 2023 international report show estimates of the difference between a country estimate and the ICILS 2023 average estimate. It is easy to see that when a country c does not contribute to the average, no sampling covariance exists between the sampling variance of the country-level estimate and the sampling variance of the average. In this case, the standard error of the difference can be expressed as the sum of both sampling variances. That is,

$$SE_{(c-\alpha)} = \sqrt{SE_c^2 + SE_{\alpha}^2}, \quad (13.19)$$

where SE_c corresponds to the standard error of the estimate of country c . The countries and benchmarking participant not included in the ICILS average are Romania, the United States, and North Rhine-Westphalia (Germany).³¹

Differences with countries included in the average

However, whenever country c did contribute to the international average, then we had to account for the fact that a sampling covariance between the country-specific estimate and the international estimate exists. In such context, the sampling covariance between both estimates equals $\frac{SE_c^2}{C^2}$. Thus, the standard error of the difference $\alpha - c$ can be expressed as,

$$\begin{aligned} SE_{(\alpha-c)} &= \sqrt{SE_c^2 + SE_\alpha^2 - \frac{2}{C^2} SE_c^2} \\ &= \sqrt{SE_\alpha^2 + \left(\frac{(C-1)^2 - 1}{C^2} \right) SE_c^2}, \end{aligned} \quad (13.20)$$

where C again corresponds to the total number of countries considered in the ICILS 2023 average, of note, the school sample of the benchmarking participant, North Rhine-Westphalia, was an explicit stratum of the German national school sample. Therefore, the estimation of the standard error of an estimated difference between North Rhine-Westphalia and an international average must take this into account. In most contexts, we used the Equation (13.20), where the sampling covariance between the international average and the estimate for North Rhine-Westphalia is accounted for. However, if a note in the table documents that Germany was excluded from the international average, Equation (13.19) was used.

Differences between CIL and CT results across 2013, 2018, and 2023

Chapter 5 of the ICILS 2023 international report includes estimates of differences in means of CIL and CT achievement over time (i.e., across 2013, 2018, and 2023). Because the process of equating the CIL and CT scores across the cycles introduced some additional error in the calculation of any test statistic, we added an equating (or linking) error term to the formula for the standard error of the difference between country averages (for more details on linking error see Chapter 11). The equating errors for CIL and CT estimates across cycles are presented in Table 13.5.

Table 13.5: Equating error for CIL and CT achievement estimates between cycles

Cycle	CIL	CT
2023-2018	2.839	2.569
2018-2013	3.901	

To incorporate the equating error in the standard error of the estimated difference across time, for example, mean difference between 2023 and 2018 in a given country ($SE_{(\theta_{2023}-\theta_{2018})}$), the squared equating error is added to formula (13.11), therefore:

$$SE_{(\theta_{2023}-\theta_{2018})} = \sqrt{SE_{\theta_{2023}}^2 + SE_{\theta_{2018}}^2 + EqErr_{(\theta_{2023}-\theta_{2018})}^2}, \quad (13.21)$$

where $SE_{\theta_{2023}}$ and $SE_{\theta_{2018}}$ corresponds to the standard error of the estimate in 2023 and 2018, respectively. In addition, $EqErr_{(\theta_{2023}-\theta_{2018})}$ corresponds to the equating error between 2023 and 2018 in CIL or CT. These equating errors are shown in Table 13.5, and correspond to 2.84 and 2.57,

31 North Rhine-Westphalia is never included in the average as a separate entity but contributes to the average as part of the German national sample.



respectively. Here, θ can be any statistic in units on the equated ICILS scale (mean, percentile, gender difference, but not percentages).

Similarly, when we compared mean achievement in a given country between 2023 and 2013, we had to incorporate the equating error of 2018–2013 ($EqErr_{(\theta_{2018}-\theta_{2013})}$) in the formula described in (13.21). This equating error was 3.90 for CIL (there is no equating error for CT between 2018 and 2013, as CT was not measured in 2013). Therefore, the standard error of the difference of CIL between 2023 and 2013, $SE_{(\theta_{2023}-\theta_{2018})}$, is expressed by:

$$SE_{(\theta_{2023}-\theta_{2018})} = \sqrt{SE_{\theta_{2023}}^2 + SE_{\theta_{2013}}^2 + EqErr_{(\theta_{2023}-\theta_{2018})}^2 + EqErr_{(\theta_{2018}-\theta_{2013})}^2}. \quad (13.22)$$

Comparison between CIL proficiency levels in 2013, 2018, and 2023

Chapter 5 of the ICILS 2023 international report also includes the difference in the percentage of students that reached proficiency Level 2 or above across 2023, 2018, and 2013. As it is impossible to directly add the equating error indicated above in such estimates, we needed to follow the replication method explained below to calculate the equating error by country.

Table 13.6: Estimates of equating error for percentages of students at or above CIL Level 2

Country	Previous cycles	Equating error 2023–2018 at or above CIL Level 2	Equating error 2018–2013 at or above CIL Level 2
Croatia	2013	0.96	*
Czech Republic	2013	1.27	*
Denmark	2018, 2013	1.13	1.25
Finland	2018	1.19	
France	2018	1.35	
Germany	2018, 2013	1.24	1.47
Italy	2018	1.52	
Kazakhstan	2018	0.80	
Korea, Republic of	2018, 2013	1.00	1.19
Luxembourg	2018	1.22	
North Rhine-W. (Germany)	2018	1.19	
Norway (Grade 9)	2013	1.36	*
Portugal	2018	1.26	
Slovak Republic	2013	1.38	*
Slovenia	2013	1.41	*
United States	2018	1.06	
Uruguay	2018	1.29	
Pooled sample		1.18 ¹	1.56 ²

Notes:

¹ Pooled equating error in 2023 includes these countries: Austria, Azerbaijan, Belgium (Flemish), Bosnia and Herzegovina, Chinese Taipei, Croatia, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Italy, Kazakhstan, Korea, Republic of, Kosovo, Latvia, Luxembourg, Malta, Norway, Oman, Portugal, Serbia, Slovak Republic, Slovenia, Spain, Sweden, and Uruguay.

² Pooled equation error in 2018 includes these countries: Chile, Denmark, Finland, France, Germany, Kazakhstan, Korea, Republic of, Luxembourg, Portugal, and Uruguay.

* Country is imputed the ICILS 2018 pooled sample equating error.

To estimate the standard error of the percentage at or above the cut-off point that defines the threshold between Levels 1 and 2 (492), within each participating country a number of n replicate cut-off points were generated by computing the observed threshold plus a random error component with a mean of 0 and a standard deviation equal to the estimated equating error (see Table 13.5). We added this component to the CIL/CT scores, replicating this procedure N times ($N = 5000$). If $EqErr_{\rho,c}$ is the equating error for proportion ρ in country c , it can be calculated as:



$$EqErr_{\rho,c} = \sqrt{\frac{\sum_{i=1}^N (\rho_{i,c} - \rho_{0,c})^2}{N}}, \quad (13.23)$$

where $\rho_{i,c}$ corresponds to the proportion of students that reached Level 2 in the replication i in country c , while $\rho_{0,c}$ corresponds to the proportion of students that reached Level 2 with the original score.

To calculate the difference in the proportion of students that reached Level 2 between 2023 and 2018, the equating error computed in (13.23) is added to the standard error of the estimated proportion of students that reached Level 2 in 2023 and 2018, analog to (13.21):

$$SE_{(\rho,c_{2023}-\rho,c_{2018})} = \sqrt{SE_{\rho,c_{2023}}^2 + SE_{\rho,c_{2018}}^2 + EqErr_{(\rho,c_{2023}-\rho,c_{2018})}^2}. \quad (13.24)$$

To estimate the difference in the proportion of students that reached Level 2 between 2023 and 2013, the equating error of 2018–2013 is added to the equating error of 2023–2018, analog to (13.22):

$$SE_{(\rho,c_{2023}-\rho,c_{2013})} = \sqrt{SE_{\rho,c_{2023}}^2 + SE_{\rho,c_{2013}}^2 + EqErr_{(\rho,c_{2023}-\rho,c_{2018})}^2 + EqErr_{(\rho,c_{2018}-\rho,c_{2013})}^2}. \quad (13.25)$$

There are some countries that participated in 2023 and 2013, but did not participate in 2018: Croatia, Czech Republic, Norway (Grade 9), Slovak Republic, and Slovenia. Since in these countries it is not possible to estimate the equating error from 2018 with 2013, their equating errors are imputed based on the pooled sample of ICILS 2018. The estimated equating errors per country and pooled sample are available in Table 13.6. Please note that the indicated equating errors are valid only to calculate the difference in the proportion of students reaching Level 2.

13.5 ICILS 2023 reporting policy

The context in which the data was generated plays an important role in understanding and interpreting the estimates reported in the ICILS 2023 international report. For this reason, in reporting ICILS 2023 country results, the tables are accompanied by the known information considered relevant for their interpretation. This ensures transparency and may highlight any limitations in their interpretation.

A first set of annotations related to a whole country. In terms of sampling, the relationship between the original sample and the realized sample (as documented by the participation rate) is relevant for interpreting and generalizing national results, as are exclusions from the target population. Countries that did not meet sampling participation requirements (combined participation rate lower than 75%), with combined school and student (or teacher) exclusion rates higher than five percent, were annotated accordingly. Other annotations included a deviation in the timing of ICILS administration and lower coverage of the nationally defined population. These annotations described important characteristics of the samples and were included in each table. A detailed description of these notes is included in Chapter 7.

Further, special attention was given to item response rates. In the ICILS 2023 international report, estimates based on a reduced weighted item response rate were annotated as follows:

- r indicates data are available for at least 70 percent but less than 85 percent of the <population>,
- s indicates data are available for at least 50 percent but less than 70 percent of the <population>,
- x indicates data are available for at least 40 percent but less than 50 percent of the <population> – to be interpreted with caution,
- y indicates data are available for less than 40 percent of the <population>, and the estimate is not reported.



Specific rules were applied for reporting or suppressing population estimates. Estimates were suppressed if participants from five or fewer jackknife zones contributed to them. In addition, when estimating the mean of a variable of interest for a subgroup, the subgroup had to consist of at least 30 participants or the estimate was suppressed.

Problems with the data collection for the ICILS main survey in the Netherlands resulted in low participation rates for all populations. As a consequence, data collected from schools and students in the Netherlands are not included in the main tables of the ICILS 2023 international report. Selected data from the Netherlands are included as an appendix and should be interpreted cautiously.

Due to problems with the data collection for the ICILS 2023 main survey in Chile, data from Chilean schools and students are not included in the ICILS 2023 international report, except data from the national context survey in Chapter 2.

13.6 Notes about software

Different software can be used to estimate population parameters with ICILS 2023 data. For the international report, we used custom codes in R in combination with the IEA IDB Analyzer for R, which is publicly available on the IEA website.³² The analysis results were verified using the IEA IDB Analyzer for SPSS to ensure consistency with previous cycles.

The IDB Analyzer is a tool for analyzing data from IEA studies. It is freely available for use with SPSS, SAS, and R. The software supports the correct estimation of population parameters and the methodology for estimating the sampling and imputation variance for most common analyses accounting for the ICILS survey design. The IDB Analyzer implements the formulas described in this chapter.

Other software supports the analysis of IEA large-scale assessment data as well. MPlus supports jackknife replication and the use of multiple plausible values. Stata supports analysis with ICILS using the `svy` command for estimation of the sampling variance and the combination of plausible values using `mi`. In addition to the IDB Analyzer, there are several other packages available in R, for example: *survey*, *RALSA*, *instvy*, or *EdSurvey* (all available in CRAN repository, the latter three are specially designed to analyze data from large-scale assessments including IEA studies).

Please note that ICILS's sampling procedures and analysis specifications can differ from those of other international large-scale assessments and even from those of other IEA studies (e.g., PIRLS, TIMSS, and ICCS). Therefore, it is recommended that you familiarize yourself with the specifics of ICILS documentation if you would like to replicate the results in the ICILS 2023 international report.

References

- Gonzalez, E., & Foy, P. (2000). Estimation of sampling variance. In M. Martin, K. Gregory, & S. Semler (Eds.), *TIMSS 1999: Technical report*. Boston College. https://timss.bc.edu/timss1999i/tech_report.html
- Jewsbury, P. A., Jia, Y., & Gonzalez, E. J. (2024). Considerations for the use of plausible values in large-scale assessments. *Large-scale Assessments in Education*, 12(1). <https://doi.org/10.1186/s40536-024-00213-y>
- Mislevy, R. J., Beaton, A., Kaplan, B. A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161. <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>
- Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(3), 355–375. <https://doi.org/10.1214/aoms/1177729989>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
- Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29(2), 614. <https://doi.org/10.1214/aoms/1177706647>

32 IEA IDB Analyzer is available at www.iea.nl/data-tools/tools [2024/11/08].



- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful?
IERI monograph series, 2(1), 9–36.
- Wolter, K. (1985). *Introduction to variance estimation*. Springer.

Appendix A:

Organizations and individuals involved in ICILS 2023

International study center

The International Study Center (ISC) is located at the International Association for the Evaluation of Educational Achievement (IEA). Center staff at IEA are responsible for designing and implementing the study in close cooperation with the National Research Coordinators (NRCs) in ICILS 2023 participating countries.

IEA is also responsible for coordinating and implementing ICILS. IEA Amsterdam, the Netherlands, is responsible for membership, translation verification, quality control monitoring, and publication. IEA Hamburg, Germany is mainly responsible for field operations, sampling procedures, and data-processing, scaling, analysis, and reporting.

Staff at IEA Amsterdam

Julian Fraillon, *international study director*
Dirk Hastedt, *executive director IEA*
Andrea Netten, *director IEA Amsterdam*
Jan-Peter Broek, *financial director IEA Amsterdam*
Isabelle Gémin, *senior financial officer*
Daniel Duckworth, *lead researcher - test development (project team)*
Lauren Musu, *head of TIQ (project team)*
Marta Moreno Hidalgo, *research officer (project team)*
David Ebbs, *senior research officer (project team)*
Kateřina Hartmanová, *senior research officer (project team)*
Katie Zuber, *head of communications*
Philippa Elliott, *publications manager*
Angela Colley, *junior publications copyeditor*
Colm Brennan, *media and communications liaison*
Morgan Kramm, *events and communications officer*

Staff at IEA Hamburg

Juliane Hencke, *director*
Sabine Meinck, *head of sampling unit and co-head of research and analysis unit*
Rolf Strietholt, *co-head of research and analysis unit*
Meng Xue, *head of software unit*

Sebastian Meyer, *ICILS co-international data manager*
Tim Daniel, *ICILS co-international data manager*

Sabine Tieck, *section lead of sampling unit*
Maximiliano Romero, *research analyst (sampling)*
Umut Atasever, *research analyst (sampling)*
Karsten Penon, *research analyst (sampling)*
Diego Cortes, *senior sampling statistician*



Duygu Savaşçı-Smith, *research analyst (sampling)*
 Rodrigo Leyton, *research analyst (sampling)*
 Pablo Torres Irribarra, *research analyst (sampling)*

Mojca Rožman, *senior research analyst (scaling and analysis coordinator)*
 Yuan-Ling Liaw, *senior research analyst (scaling)*
 Andrés Christiansen, *senior research analyst (scaling)*
 Minge Chen, *research analyst (scaling)*
 Ana María Mejía-Rodríguez, *senior research analyst (scaling support)*
 Andrés Strello, *senior research analyst (analysis and table design)*
 Marlen Holtmann, *senior research analyst (analysis)*
 Ann-Kristin Koop, *research analyst (analysis verification)*
 Nurullah Eryilmaz, *senior research analyst (analysis verification)*
 Falk Brese, *senior research analyst (analysis verification)*
 Alec Kennedy, *senior research analyst (communications support)*

Farhana Tabassum, *study coordinator 2024 onwards*
 Zarrinigor Nozimova, *study coordinator 2023*
 Eleonora Kolomiets, *study coordinator 2022–2023*
 Juliane Kobelt, *study coordinator 2020–2022*

Svetoslav Velkov, *tester (software testing)*
 Vanisa Vongphanakhone, *research analyst (national adaptations and material receipt)*
 Svenja Kalmbach, *research analyst (national adaptations and data processing)*
 Valentina Rivera Toloza, *research analyst (national adaptations)*
 Isbat Ibn Hasnat, *research analyst (software requirements and testing)*
 Delnaz Mohebigilani, *programmer (software development)*
 Limiao Duan, *programmer (software development)*
 Abdullah Al-Azzawi, *research analyst (data processing systems)*
 Jurij Lenar, *research analyst (meta-data and codebooks)*
 Renato Alves Coppi, *research analyst (data processing)*
 Widiyanto Persadha, *research analyst (data processing systems)*
 Daniel Ugurel, *research analyst (online questionnaires and material receipt)*
 Mohamadreza Atrian, *programmer (software development)*
 Wolfram Jarchow, *research analyst (meta-data and codebooks)*
 Ankita Kulkarni, *research analyst (software requirements and testing)*
 Miguel Hernandez Acosta, *programmer (software development)*
 Ahmad Khalil, *research analyst (sampling, player testing, and layout verification)*
 Yagmur Rizvi, *research analyst (software testing)*
 Xiao Sui, *research analyst (software requirements and testing)*
 Chen-Ho Lee, *research analyst (material receipt)*
 Oriana Mora, *research analyst (national adaptations and codebooks)*
 Christine Busch, *research analyst (meta-data and quality control)*
 Alischa Dietz, *research analyst (data processing and derived variables)*
 Laiza Souza, *event coordinator*
 Aisha Elsayed, *student assistant*
 Osman Deger, *student assistant*
 Adham Kotb, *student assistant*
 Inderawes Khalil, *student assistant*
 Tam Nguyen, *student assistant*
 Duygu Uyar, *student assistant*



Marwan Elkholy, *student assistant*
 Hajar Zare, *student assistant*
 Ahmed Aboelela, *student assistant*

LaTeX design support

Velimir Gayevskiy, *Creodocs Ltd., Boston, USA*

RM Assessment

RM Assessment was responsible for developing the software systems underpinning the computer-based student assessment instruments for ICILS 2023. This work included development of the test and questionnaire items, the assessment delivery system, and the web-based translation, scoring, and data-management modules.

Staff RM Assessment

Rajani de Man, *customer relationship manager*
 Neal Varghese, *project manager*
 Monica Srinivasan, *project manager*
 Derek van Neuren, *development manager*
 David Wynn, *development technical lead*
 Felicia Gu, *development technical lead*
 Ian Qiu, *software developer*
 Tinu Treesa Tom, *software developer*
 Reshma Merin, *software developer*
 Reeya Grace Philip, *software developer*
 Shagna Cheerngodan, *quality assurance*
 Geethu Parameswaran Nair, *quality assurance*
 Anjana Albert, *quality assurance*
 Stephen Ainley, *customer support*
 Shawn Habibnia, *customer support*
 Darren Blakely, *customer support*
 Ranil Weerasinghe, *customer support*
 Jim Murdoch, *customer support*
 Rohit Koshy Kurian, *senior system analyst*
 Rojan Jojo, *senior system analyst*

ICILS sampling referee

Marc Joncas was the sampling referee for the study. He has provided invaluable advice on all sampling-related aspects of the study.

National research coordinators

The national research coordinators (NRCs) played a crucial role in the development of the project. They provided policy- and content-oriented advice on the development of the instruments and were responsible for the implementation of ICILS in the participating countries.

Austria

Iris Höller
 Magdalena Rölz
Federal Institute for Quality Assurance of the Austrian School System (IQS)

Azerbaijan

Shirin Mammadov
The Institute of Education of the Republic of Azerbaijan

**Belgium (Flemish)**

Charlotte Struyve

*Centre for Educational Effectiveness and Evaluation, KU Leuven***Bosnia and Herzegovina**

Nataša Kokoruš

*Agency for Pre-Primary, Primary and Secondary Education***Chile**

Millicent Bader

Gabriel Mendoza Gonzalez 2024 onwards

*National Agency for Educational Quality***Chinese Taipei**

Cheng-Chih Wu

Meng-Jung Tsai

*School of Learning Informatics, National Taiwan Normal University***Croatia**

Hrvoje Mladinić

*Nacionalni centar za vanjsko vrednovanje obrazovanja***Cyprus**

Yiasemina Karagiorgi

*Cyprus Pedagogical Institute, Cyprus Ministry of Education, Sport and Youth***Czech Republic**

Josef Basl

*Czech School Inspectorate***Denmark**

Jeppe Bundsgaard

*Danish School of Education, Aarhus University***Finland**

Kaisa Leino

Janne Fagerlund

*University of Jyväskylä, Finnish Institute for Educational Research***France**

Elodie Persem

Laurent Korchia 2025 onwards

Julie Thumerelle 2018–2024

*Ministry of National Education***Germany**

Birgit Eickelmann

*Paderborn University***Germany, North-Rhine Westphalia**

Kerstin Drossel

*Paderborn University***Greece**

Maria Nika

Institute of Educational Policy (IEP)

Hungary

Andrea Fási 2024 onwards

Annamária Rábainé Szabó 2018–2024

Educational Authority

Italy

Carlo Di Chiacchio

Elisa Caponera

INVALSI

Kazakhstan

Munira Vassilova 2023 onwards

Nazym Smanova 2022–2023

Aigul Baigulova 2022

Aiym Chegenbayeva 2021

National Center for Education Research and Evaluation Taldau named after Ahmet Baitursynuly

Kosovo

Besa Gashi

Birsen Sufta

Ministry of Education, Science, Technology and Innovation

Latvia

Olga Pole

University of Latvia

Luxembourg

Catalina Lomos

Luxembourg Institute of Socio-Economic Research and

Elisa Mazzucato

Ministry of Education, Children and Youth, Luxembourg Institute of Socio-Economic Research

Malta

Jude Zammit

Ministry for Education, Sport, Youth, Research and Innovation

The Netherlands

Alexander Krepel

Kohnstamm Instituut

Norway

Anubha Rohatgi

Department of Teacher Education & School Research

Oman

Maya Said Al Azri 2023 onwards

Zuwaina Saleh Almaskari 2018–2023

Ministry of Education

Portugal

Anabela Serrão

Institute for Educational Evaluation (IAVE, I.P.)

Republic of Korea

Hyunwoo Kim 2024 onwards

Minhee Seo 2021–2024

Seongkyun Jeon 2018–2021



Korea Institute for Curriculum and Evaluation

Romania

Olimpiu Istrate 2024 onwards

Dragos Iliescu

University of Bucharest

Serbia

Katarina Aleksić

Education Technology Center, Institute for Education Quality and Evaluation

Slovak Republic

Romana Panáčková

National Institute of Education and Youth (NIVaM)

Slovenia

Eva Klemenčič Mirazchiyski

Educational Research Institute

Spain

Ricardo Callejo

National Institute for Educational Assessment, Ministry of Education

Sweden

Fredrik Aspling

Swedish National Agency for Education

United States

Linda Hamilton

National Center for Education Statistics

Uruguay

Valentina Cencela 2024 onwards

Mariana González 2024 onwards

Cecilia Hughes 2018–2024

Ceibal

Appendix B:

National characteristics

This appendix describes, for each education system participating in ICILS 2023, the population coverage, exclusion categories, stratification variables, and any deviations from the general ICILS 2023 sampling design.

The same sample of schools was selected for the student survey and the teacher survey. However, the school participation status of a school in the student and teacher survey could differ. It was particularly common for a school to count as participating in the student survey, but not in the teacher survey; however, the reverse scenario was also possible.

B.1 Austria

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than three students in the target grade) and foreign schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school type (APS—Allgemeine Pflichtschule, AHS—Allgemeinbildende Höhere Schule, Miscellaneous School Types), ISF—Individual Student Funding (high, low), and urbanization (rural, urban), resulting in seven explicit strata.
- Implicit stratification was applied by states, giving a total of nine implicit strata.
- The sample was disproportionally allocated to explicit strata, due to the minimum sample size requirement of very small strata.
- Small schools were selected with equal probabilities.
- The main survey samples for ICILS 2023 and TIMSS 2023 were selected simultaneously, resulting in a minimum overlap of the sampled schools.

Table B.1: Allocation of student sample in Austria

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
AHS with Low ISF	40	0	40	0	0	0	0
AHS with High ISF	14	0	14	0	0	0	0
Rural APS with Low ISF	28	0	27	0	0	1	0
Rural APS with High ISF	10	0	10	0	0	0	0
Urban APS with Low ISF	36	1	35	0	0	0	0
Urban APS with High ISF	24	0	24	0	0	0	0
Miscellaneous School Type	8	2	2	2	0	0	2
Total	160	3	152	2	0	1	2

Note: No schools with student participation rate below 50% were found.



Table B.2: Allocation of teacher sample in Austria

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
AHS with Low ISF	40	0	35	0	0	5	0
AHS with High ISF	14	0	12	0	0	2	0
Rural APS with Low ISF	28	0	26	0	0	2	0
Rural APS with High ISF	10	0	9	0	0	1	0
Urban APS with Low ISF	36	1	29	0	0	6	0
Urban APS with High ISF	24	0	18	0	0	6	0
Miscellaneous School Type	8	2	2	2	0	0	2
Total	160	3	131	2	0	22	2

Note: Twenty-two schools were regarded non-participating as the within-school participation rate was below 50%.

B.2 Azerbaijan

- School-level exclusions consisted of very small schools (fewer than five students in the target grade), school with language of instruction other than Azerbaijani, and schools for children with special educational needs. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by language (Azerbaijani only, Azerbaijani and Russian), and urbanization (urban in main cities, urban in other regions, rural in main cities, rural in other regions), resulting in five explicit strata.
- Implicit stratification was applied for Azerbaijani and Russian schools by urbanization (urban in main cities, urban in other regions, rural in main cities, rural in other regions), giving a total of four implicit strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve maximum overlap between these studies. Different classes were randomly selected for both studies.

Table B.3: Allocation of student sample in Azerbaijan

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Azerbaijani Only - Urban (Main Cities)	32	0	30	2	0	0	0
Azerbaijani Only - Urban (Other Regions)	18	0	15	3	0	0	0
Azerbaijani Only - Rural (Main Cities)	10	0	10	0	0	0	0
Azerbaijani Only - Rural (Other Regions)	62	0	59	2	1	0	0
Azerbaijani and Russian	42	0	41	1	0	0	0
Total	164	0	155	8	1	0	0

Note: No schools with student participation rate below 50% were found.

Table B.4: Allocation of teacher sample in Azerbaijan

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Azerbaijani Only - Urban (Main Cities)	32	0	28	2	0	2	0
Azerbaijani Only - Urban (Other Regions)	18	0	12	3	0	3	0
Azerbaijani Only - Rural (Main Cities)	10	0	10	0	0	0	0
Azerbaijani Only - Rural (Other Regions)	62	0	51	2	1	8	0
Azerbaijani and Russian	42	0	37	1	0	4	0
Total	164	0	138	8	1	17	0

Note: Seventeen schools were regarded non-participating as the within-school participation rate was below 50%.

B.3 Belgium (Flemish)

- School-level exclusions consisted of schools for children with special educational needs and very small schools (fewer than ten students in the target grade). Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school type (Algemeen Secundair/Kunst Secundair (ASO-/KSO), Technisch Secundair/Beroeps Secundair (TSO/BSO), Multilateral/Middenschool Autonomoom (Multilat./AutoMid), included SEN schools), resulting in four explicit strata.



- Implicit stratification was applied by school administration type (public, private), giving a total of two implicit strata.
- Small schools were selected with equal probabilities.

Table B.5: Allocation of student sample in Belgium (Flemish)

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
ASO/KSO	40	0	25	9	1	5	0
TSO/BSO	28	0	21	3	0	4	0
Multilat./ AutoMid	78	1	48	20	3	6	0
Included SEN School	8	0	2	4	0	2	0
Total	154	1	96	36	4	17	0

Note: No schools with student participation rate below 50% were found.

Table B.6: Allocation of teacher sample in Belgium (Flemish)

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
ASO/KSO	40	0	24	8	1	7	0
TSO/BSO	28	0	18	3	0	7	0
Multilat./AutoMid	78	1	43	19	3	12	0
Included SEN School	8	0	1	4	0	3	0
Total	154	1	86	34	4	29	0

Note: Twelve schools were regarded non-participating as the within-school participation rate was below 50%.

B.4 Bosnia and Herzegovina

- Bosnia and Herzegovina participated without Republika Srpska resulting in a reduced coverage of 60.2%.
- School-level exclusions consisted of schools in West Herzegovina Canton, schools for children with special educational needs, very small schools (fewer than five students in the target grade), and international schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by canton, resulting in ten explicit strata.
- No implicit stratification was applied.
- The sample was disproportionately allocated to explicit strata, due to the minimum sample size requirement of very small strata.



- Small schools were not selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve maximum overlap between these studies. Different classes were randomly selected for both studies.

Table B.7: Allocation of student sample in Bosnia and Herzegovina

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Bosnian-Podrinje Canton Goražde	6	0	6	0	0	0	0
Brčko District	5	0	5	0	0	0	0
Sarajevo Canton	20	0	20	0	0	0	0
Canton 10	6	0	6	0	0	0	0
Herzegovina-Neretva Canton	9	0	8	1	0	0	0
Posavina Canton	6	0	6	0	0	0	0
Central Bosnia Canton	9	0	7	1	0	1	0
Tuzla Canton	21	0	21	0	0	0	0
Una-Sana Canton	9	0	9	0	0	0	0
Zenica-Doboj Canton	15	0	15	0	0	0	0
Total	106	0	103	2	0	1	0

Note: One school was regarded non-participating as the within-school participation rate was below 50%.

Table B.8: Allocation of teacher sample in Bosnia and Herzegovina

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
All	106	0	37	1	0	68	0
Total	106	0	37	1	0	68	0

Note: Sixty-four schools were regarded non-participating as the within-school participation rate was below 50%. Original strata were collapsed as three strata had no participating schools.

B.5 Chile

- School-level exclusions consisted of very small schools (fewer than five students in the target grade) geographically inaccessible schools, and schools for children with special educational need. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.



- Explicit stratification was performed by school administration type (public, subsidized schools, private), resulting in three explicit strata.
- Implicit stratification was applied by the national assessment performance group for mathematics, giving a total of four implicit strata.
- The sample was disproportionately allocated to explicit strata to allow comparison.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS main survey sample was selected by controlling for the overlap with the TIMSS Main Data Collection sample using the Chowdhury approach.

Table B.9: Allocation of student sample in Chile

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Public schools	60	0	47	2	0	11	0
Subsidized schools	78	0	60	2	0	16	0
Private schools	30	0	16	2	0	12	0
Total	168	0	123	6	0	39	0

Note: Five schools were regarded non-participating as the within-school participation rate was below 50%.

Table B.10: Allocation of teacher sample in Chile

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Public schools	60	0	21	0	0	39	0
Subsidized schools	78	0	33	1	0	44	0
Private schools	30	0	3	2	0	25	0
Total	168	0	57	3	0	108	0

Note: Sixty-nine schools were regarded non-participating as the within-school participation rate was below 50%.

B.6 Chinese Taipei

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade) and schools with a different curriculum. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by urbanization (metropolitan area, city area, town area, rural area, remote area) and region (north, other), resulting in eight explicit strata.



- Implicit stratification was applied by school performance, giving a total of five implicit strata.
- The sample was disproportionally allocated to explicit strata, due to the minimum sample size requirement of very small strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.

Table B.11: Allocation of student sample in Chinese Taipei

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Metropolitan Area - North	21	0	21	0	0	0	0
Metropolitan Area - Others	9	0	9	0	0	0	0
City Area - North	24	0	23	1	0	0	0
City Area - Others	20	0	20	0	0	0	0
Town Area - North	20	0	20	0	0	0	0
Town Area - Others	35	0	34	1	0	0	0
Rural Area - All	19	0	18	1	0	0	0
Remote Area - All	21	0	20	1	0	0	0
Total	169	0	165	4	0	0	0

Note: No schools with student participation rate below 50% were found.

Table B.12: Allocation of teacher sample in Chinese Taipei

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Metropolitan Area - North	21	0	20	0	0	1	0
Metropolitan Area - Others	9	0	9	0	0	0	0
City Area - North	24	0	23	1	0	0	0
City Area - Others	20	0	20	0	0	0	0
Town Area - North	20	0	20	0	0	0	0
Town Area - Others	35	0	33	1	0	1	0
Rural Area - All	19	0	18	1	0	0	0
Remote Area - All	21	0	20	1	0	0	0
Total	169	0	163	4	0	2	0

Note: No schools with teacher participation rate below 50% were found.



B.7 Croatia

- School-level exclusions consisted of schools for children with special educational needs and Minority Schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by region (Central Croatia, Eastern Croatia, Northern Croatia, Western Croatia, Southern Croatia, City of Zagreb), resulting in six explicit strata.
- No implicit stratification was applied.
- Small schools were selected with equal probabilities.

Table B.13: Allocation of student sample in Croatia

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	
	Ineligible Schools	Sampled Schools	First Replacements	Second Replacements	Excluded Schools		
Central Croatia	32	0	30	2	0	0	0
Eastern Croatia	24	0	20	4	0	0	0
Northern Croatia	16	0	15	1	0	0	0
Western Croatia	18	0	15	2	1	0	0
Southern Croatia	32	0	25	5	0	2	0
City of Zagreb	30	0	22	6	0	2	0
Total	152	0	127	20	1	4	0

Note: Three schools were regarded non-participating as the within-school participation rate was below 50%.

Table B.14: Allocation of teacher sample in Croatia

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	
	Ineligible Schools	Sampled Schools	First Replacements	Second Replacements	Excluded Schools		
Central Croatia	32	0	29	2	0	1	0
Eastern Croatia	24	0	20	3	0	1	0
Northern Croatia	16	0	15	1	0	0	0
Western Croatia	18	0	15	2	1	0	0
Southern Croatia	32	0	25	6	0	1	0
City of Zagreb	30	0	22	6	0	2	0
Total	152	0	126	20	1	5	0

Note: Five schools were regarded non-participating as the within-school participation rate was below 50%.



B.8 Cyprus

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade), language of instruction other than Greek or English, and schools in the Turkish area. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by district (Nicosia, Limassol, Famagusta-Larnaca, Paphos), school administration (public, private), and curriculum (national, other), resulting in five explicit strata.
- Implicit stratification was applied by urbanization (urban, rural), giving a total of two implicit strata.
- Census of schools and students. All variance estimations were computed using classes as variance strata.

Table B.15: Allocation of student sample in Cyprus

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	Excluded Schools
		Ineligible Schools	Sampled Schools	First Replacements	Second Replacements		
Public or Private with national curriculum - Nicosia	29	0	29	0	0	0	0
Public or Private with national curriculum - Limassol	23	0	23	0	0	0	0
Public or Private with national curriculum - Famagusta-Larnaca	16	0	16	0	0	0	0
Public or Private with national curriculum - Paphos	8	0	8	0	0	0	0
Private with other curriculum - All	23	0	19	0	0	4	0
Total	99	0	95	0	0	4	0

Note: No schools with student participation rate below 50% were found.



Table B.16: Allocation of teacher sample in Cyprus

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Public or Private with national curriculum - Nicosia	29	0	29	0	0	0	0
Public or Private with national curriculum - Limassol	23	0	22	0	0	1	0
Public or Private with national curriculum - Famagusta-Larnaca	16	0	14	0	0	2	0
Public or Private with national curriculum - Paphos	8	0	8	0	0	0	0
Private with other curriculum - All	23	0	16	0	0	7	0
Total	99	0	89	0	0	10	0

Note: Six schools were regarded non-participating as the within-school participation rate was below 50%.

B.9 Czech Republic

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than four students in the target grade), and Polish language schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school type (basic school, multi-year gymnasium), and district (Praha, Středočeský, Jihočeský, Plzeňský, Karlovarský, Ústecký, Liberecký, Královéhradecký, Pardubický, Vysočina, Jihomoravský, Olomoucký, Zlínský, Moravskoslezský), resulting in 15 explicit strata.
- Implicit stratification by district was applied only for multi-year gymnasium, giving a total of 14 implicit strata.
- The sample was disproportionately allocated to explicit strata.
- Small schools were selected with equal probabilities.
- The main survey samples for ICILS 2023 and TIMSS 2023 were selected simultaneously, resulting in a minimum overlap of the sampled schools.



Table B.17: Allocation of student sample in Czech Republic

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Basic School – Praha	16	0	16	0	0	0	0
Basic School - Středočeský	20	0	20	0	0	0	0
Basic School - Jihočeský	10	0	10	0	0	0	0
Basic School - Plzeňský	8	0	8	0	0	0	0
Basic School - Karlovarský	12	0	12	0	0	0	0
Basic School - Ústecký	20	0	20	0	0	0	0
Basic School - Liberecký	8	0	8	0	0	0	0
Basic School - Královéhradecký	8	0	8	0	0	0	0
Basic School - Pardubický	8	0	8	0	0	0	0
Basic School - Vysočina	8	0	8	0	0	0	0
Basic School - Jihomoravský	24	0	24	0	0	0	0
Basic School - Olomoucký	8	0	8	0	0	0	0
Basic School - Zlínský	8	0	8	0	0	0	0
Basic School - Moravskoslezský	24	0	24	0	0	0	0
Multi-year Gymnasium - All	28	0	28	0	0	0	0
Total	210	0	210	0	0	0	0

Note: No schools with student participation rate below 50% were found.



Table B.18: Allocation of teacher sample in Czech Republic

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Basic School - Praha	16	0	16	0	0	0	0
Basic School - Středočeský	20	0	20	0	0	0	0
Basic School - Jihočeský	10	0	10	0	0	0	0
Basic School - Plzeňský	8	0	8	0	0	0	0
Basic School - Karlovarský	12	0	12	0	0	0	0
Basic School - Ústecký	20	0	20	0	0	0	0
Basic School - Liberecký	8	0	8	0	0	0	0
Basic School - Královéhradecký	8	0	8	0	0	0	0
Basic School - Pardubický	8	0	8	0	0	0	0
Basic School - Vysočina	8	0	8	0	0	0	0
Basic School - Jihomoravský	24	0	24	0	0	0	0
Basic School - Olomoucký	8	0	8	0	0	0	0
Basic School - Zlínský	8	0	8	0	0	0	0
Basic School - puMoravskoslezský	24	0	24	0	0	0	0
Multi-year Gymnasium - All	28	0	28	0	0	0	0
Total	210	0	210	0	0	0	0

Note: No schools with teacher participation rate below 50% were found.

B.10 Denmark

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade), and schools teaching in other language than Danish. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- No explicit stratification was applied.
- Implicit stratification was applied by achievement score (grade point average), giving a total of five implicit strata.
- Small schools were selected with equal probabilities.



Table B.19: Allocation of student sample in Denmark

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
All	150	2	118	15	8	7	0
Total	150	2	118	15	8	7	0

Note: No schools with student participation rate below 50% were found.

Table B.20: Allocation of teacher sample in Denmark

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
All	150	2	104	15	6	23	0
Total	150	2	104	15	6	23	0

Note: Fifteen schools were regarded non-participating as the within-school participation rate was below 50%.

B.11 Finland

- School-level exclusions consisted of schools for children with special educational needs and language schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school language (Finnish speaking, Swedish speaking), regional state administrative agency (Helsinki-Uusimaa, Southern, Western, Northern and Eastern) and urbanization (rural, urban and semi-urban), resulting in eight explicit strata.
- Implicit stratification was applied only for Swedish speaking schools by the regional state administrative agency, giving a total of six implicit strata.
- The sample was disproportionally allocated to explicit strata, due to the minimum sample size requirement of very small strata.
- Small schools were selected with equal probabilities.
- The main survey samples for ICILS 2023 and TIMSS 2023 were selected simultaneously, resulting in a minimum overlap of the sampled schools.



Table B.21: Allocation of student sample in Finland

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Finnish speaking - Helsinki-Uusimaa - All	42	0	42	0	0	0	0
Finnish speaking - Southern - Urban and Semi-Urban	26	0	26	0	0	0	0
Finnish speaking - Southern - Rural	8	0	8	0	0	0	0
Finnish speaking - Western - Urban and Semi-Urban	30	0	30	0	0	0	0
Finnish speaking - Western - Rural	8	0	8	0	0	0	0
Finnish speaking - Northern and Eastern - Urban and Semi-Urban	28	0	28	0	0	0	0
Finnish speaking - Northern and Eastern - Rural	8	0	8	0	0	0	0
Swedish speaking - All	8	0	8	0	0	0	0
Total	158	0	158	0	0	0	0

Note: No schools with student participation rate below 50% were found.



Table B.22: Allocation of teacher sample in Finland

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Finnish speaking - Helsinki-Uusimaa - All	42	0	40	0	0	2	0
Finnish speaking - Southern - Urban and Semi-Urban	26	0	25	0	0	1	0
Finnish speaking - Southern - Rural	8	0	8	0	0	0	0
Finnish speaking - Western - Urban and Semi-Urban	30	0	30	0	0	0	0
Finnish speaking - Western - Rural	8	0	8	0	0	0	0
Finnish speaking - Northern and Eastern - Urban and Semi-Urban	28	0	28	0	0	0	0
Finnish speaking - Northern and Eastern - Rural	8	0	8	0	0	0	0
Swedish speaking - All	8	0	8	0	0	0	0
Total	158	0	155	0	0	3	0

Note: Three schools were regarded non-participating as the within-school participation rate was below 50

B.12 France

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade), private schools without a contract, Mayotte schools, and overseas territories. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school administration type (Public Schools outside Priority Education, Public Schools in Priority Education, Private Schools), resulting in three explicit strata.
- No implicit stratification was applied.
- Small schools were selected with equal probabilities.



Table B.23: Allocation of student sample in France

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Public Schools outside Priority Education	95	0	95	0	0	0	0
Public Schools in Priority Education	24	0	24	0	0	0	0
Private Schools	33	1	31	0	0	1	0
Total	152	1	150	0	0	1	0

Note: No schools with student participation rate below 50% were found.

Table B.24: Allocation of teacher sample in France

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Public Schools outside Priority Education	95	0	77	0	0	18	0
Public Schools in Priority Education	24	0	17	0	0	7	0
Private Schools	33	1	25	0	0	7	0
Total	152	1	119	0	0	32	0

Note: Thirty-two schools were regarded non-participating as the within-school participation rate was below 50%.

B.13 Germany

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than three students in the target grade), and international schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by German state (Nordrhein-Westfalen, other states) and school type (Gymnasium, Non-Gymnasium, schools for children with special educational needs), resulting in five explicit strata.
- Implicit stratification was applied by German state and socioeconomic status (migrant proportion), giving a total of 50 implicit strata.
- The sample was disproportionately allocated to explicit strata.
- Small schools were not selected with equal probabilities.



Table B.25: Allocation of student sample in Germany

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
North Rhine-Westphalia - Gymnasium	42	0	38	3	0	1	0
North Rhine-Westphalia - All other school types	70	0	65	4	0	1	0
All other states - Gymnasium	44	0	37	5	1	1	0
All other states - All Non-Gymnasium	70	0	57	5	4	4	0
All states - SEN schools	4	0	2	1	0	1	0
Total	230	0	199	18	5	8	0

Note: Two schools were regarded non-participating as the within-school participation rate was below 50%.

Table B.26: Allocation of teacher sample in Germany

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
North Rhine-Westphalia - Gymnasium	42	0	37	3	0	2	0
North Rhine-Westphalia - All other school types	70	0	62	3	0	5	0
All other states - Gymnasium	44	0	27	3	1	13	0
All other states - All other school types	70	0	37	2	3	28	0
All states - SEN schools	4	0	1	1	0	2	0
Total	230	0	164	12	4	50	0

Note: Forty-two schools were regarded non-participating as the within-school participation rate was below 50%.

B.14 Greece

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than six students in the target grade). Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school administration (public, private) and administrative regions (Attica, Central Greece, Central Macedonia, Crete, East Macedonia and Thrace, Epirus, Ionian Islands, North Aegean, Peloponnese, South Aegean, Thessaly, West Greece, West Macedonia, Rest of Greece), resulting in 15 explicit strata.
- Implicit stratification was applied by urbanization, giving a total of five implicit strata.



- The sample was disproportionally allocated to explicit strata, due to the minimum sample size requirement of very small strata.
- Small schools were selected with equal probabilities.

Table B.27: Allocation of student sample in Greece

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Public - Attica	48	0	38	6	1	3	0
Public - Central Greece	8	0	7	0	0	1	0
Public - Central Macedonia	26	0	25	0	1	0	0
Public - Crete	12	0	10	0	0	2	0
Public - East Macedonia and Thrace	8	0	5	3	0	0	0
Public - Epirus	8	0	5	2	0	1	0
Public - Ionian Islands	8	0	7	1	0	0	0
Public - North Aegean	8	0	8	0	0	0	0
Public - Peloponnese	8	0	8	0	0	0	0
Public - South Aegean	8	0	8	0	0	0	0
Public - Thessaly	10	0	9	1	0	0	0
Public - West Greece	10	0	10	0	0	0	0
Public - West Macedonia	8	0	7	1	0	0	0
Private - Attica	8	0	7	1	0	0	0
Private - Rest of Greece	8	0	7	1	0	0	0
Total	186	0	161	16	2	7	0

Note: One school was regarded non-participating as the within-school participation rate was below 50%.



Table B.28: Allocation of teacher sample in Greece

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Public - Attica	48	0	36	5	1	6	0
Public - Central Greece	8	0	6	0	0	2	0
Public - Central Macedonia	26	0	24	0	1	1	0
Public - Crete	12	0	11	0	0	1	0
Public - East Macedonia & Thrace	8	0	5	3	0	0	0
Public - Epirus	8	0	5	2	0	1	0
Public - Ionian Islands	8	0	6	1	0	1	0
Public - North Aegean	8	0	8	0	0	0	0
Public - Peloponnese	8	0	8	0	0	0	0
Public - South Aegean	8	0	7	0	0	1	0
Public - Thessaly	10	0	9	1	0	0	0
Public - West Greece	10	0	10	0	0	0	0
Public - West Macedonia	8	0	7	1	0	0	0
Private - Attica	8	0	7	1	0	0	0
Private - Rest of Greece	8	0	7	0	0	1	0
Total	186	0	156	14	2	14	0

Note: Eleven schools were regarded non-participating as the within-school participation rate was below 50%.

B.15 Hungary

- School-level exclusions consisted of schools for children with special educational needs and foreign language schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by urbanization (capital, county town, town, rural area) and national assessment mathematics score (high, medium, low or medium, low), resulting in 10 explicit strata.
- Implicit stratification was applied by region, giving a total of seven implicit strata.
- Small schools were selected with equal probabilities.



Table B.29: Allocation of student sample in Hungary

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools		Non-participating Schools	Excluded Schools	
			Sampled Schools	First Replacements	Second Replacements		
Capital - High	8	0	6	0	1	1	0
Capital - Low or Medium	18	0	18	0	0	0	0
County town - High	10	0	10	0	0	0	0
County town - Low or Medium	22	0	21	0	1	0	0
Town - Low	22	0	22	0	0	0	0
Town - Medium	20	0	20	0	0	0	0
Town - High	20	0	20	0	0	0	0
Rural area - Low	14	1	12	0	0	1	0
Rural area - Medium	12	0	12	0	0	0	0
Rural area - High	12	0	12	0	0	0	0
Total	158	1	153	0	2	2	0

Note: No schools with student participation rate below 50% were found.

Table B.30: Allocation of teacher sample in Hungary

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools		Non-participating Schools	Excluded Schools	
			Sampled Schools	First Replacements	Second Replacements		
Capital - High	8	0	5	0	1	2	0
Capital - Low or Medium	18	0	13	0	0	5	0
County town - High	10	0	10	0	0	0	0
County town - Low or Medium	22	0	21	0	1	0	0
Town - Low	22	0	19	0	0	3	0
Town - Medium	20	0	20	0	0	0	0
Town - High	20	0	18	0	0	2	0
Rural area - Low	14	1	12	0	0	1	0
Rural area - Medium	12	0	12	0	0	0	0
Rural area - High	12	0	10	0	0	2	0
Total	158	1	140	0	2	15	0

Note: Thirteen schools were regarded non-participating as the within-school participation rate was below 50%.

B.16 Italy

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than three students in the target grade), Ladin language schools, German language schools, and Slovenian language schools. Within-school exclusions consisted of stu-



dents (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.

- Explicit stratification was performed by region (center, south islands, north east, north west, south), resulting in five explicit strata.
- Implicit stratification was applied by school administration type (two types) and national test results (three levels), giving a total of five implicit strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.

Table B.31: Allocation of student sample in Italy

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Center	30	0	28	2	0	0	0
South Islands	22	0	21	1	0	0	0
North East	30	0	29	1	0	0	0
North West	40	0	38	2	0	0	0
South	30	0	29	1	0	0	0
Total	152	0	145	7	0	0	0

Note: No schools with student participation rate below 50% were found.

Table B.32: Allocation of teacher sample in Italy

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Center	30	0	28	2	0	0	0
South Islands	22	0	21	1	0	0	0
North East	30	0	29	1	0	0	0
North West	40	0	38	2	0	0	0
South	30	0	28	1	0	1	0
Total	152	0	144	7	0	1	0

Note: One school was regarded non-participating as the within-school participation rate was below 50%.

B.17 Kazakhstan

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than six students in the target grade), minority language schools, geograph-



ically inaccessible schools, and evening schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.

- Explicit stratification was performed by region, resulting in 17 explicit strata.
- Implicit stratification was applied by urbanization (two levels) and school language (three types of language), giving a total of six implicit strata.
- The sample was disproportionately allocated to explicit strata, due to the minimum sample size requirement of very small strata.
- Small schools were selected with equal probabilities.

Table B.33: Allocation of student sample in Kazakhstan

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Akmola	8	0	8	0	0	0	0
Mangystau	8	0	8	0	0	0	0
Turkistan	22	0	21	0	0	0	1
Pavlodar	8	0	8	0	0	0	0
East	10	0	8	1	0	0	1
North	8	0	8	0	0	0	0
Nur-Sultan city	8	0	8	0	0	0	0
Almaty city	14	0	14	0	0	0	0
Shymkent city	10	0	10	0	0	0	0
Atyrau	8	0	8	0	0	0	0
Aktobe	8	0	8	0	0	0	0
Almaty	20	0	20	0	0	0	0
West	8	0	8	0	0	0	0
Karaganda	10	0	7	3	0	0	0
Kostanay	8	0	8	0	0	0	0
Zhambyl	12	0	12	0	0	0	0
Kyzylorda	8	0	8	0	0	0	0
Total	178	0	172	4	0	0	2

Note: No schools with student participation rate below 50% were found.



Table B.34: Allocation of teacher sample in Kazakhstan

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools		Second Replace-ments	Non-partici-pating Schools	Excluded Schools
			Sampled Schools	First Replace-ments			
Akmola	8	0	8	0	0	0	0
Mangystau	8	0	8	0	0	0	0
Turkistan	22	0	21	0	0	0	1
Pavlodar	8	0	8	0	0	0	0
East	10	0	8	1	0	0	1
North	8	0	8	0	0	0	0
Nur-Sultan city	8	0	8	0	0	0	0
Almaty city	14	0	14	0	0	0	0
Shymkent city	10	0	10	0	0	0	0
Atyrau	8	0	8	0	0	0	0
Aktobe	8	0	8	0	0	0	0
Almaty	20	0	20	0	0	0	0
West	8	0	8	0	0	0	0
Karaganda	10	0	7	3	0	0	0
Kostanay	8	0	8	0	0	0	0
Zhambyl	12	0	12	0	0	0	0
Kyzylorda	8	0	8	0	0	0	0
Total	178	0	172	4	0	0	2

Note: No schools with teacher participation rate below 50% were found.

B.18 Korea, Republic of

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade), alternative schools, and remote schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by urbanization (big city, medium or small city, small town or village) and school type by gender (girl, boy, mixed), resulting in nine explicit strata.
- Implicit stratification was applied by city, giving a total of 17 implicit strata.
- The sample was disproportionately allocated to explicit strata, due to the minimum sample size requirement of very small strata.
- Small schools were selected with equal probabilities.
- The main survey samples for ICILS 2023 and TIMSS 2023 were selected simultaneously, resulting in maximum overlap in the first six strata and minimum overlap in strata seven to nine. In schools selected for both studies, different classes were randomly selected for both studies.



Table B.35: Allocation of student sample in Republic of Korea

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Big City - Boy	8	0	6	2	0	0	0
Big City - Girl	8	0	6	1	1	0	0
Big City - Mixed	40	0	25	13	2	0	0
Medium or Small City - Boy	8	0	5	3	0	0	0
Medium or Small City - Girl	8	0	7	0	1	0	0
Medium or Small City - Mixed	48	0	40	6	2	0	0
Small Town or Village - Boy	8	0	6	2	0	0	0
Small Town or Village - Girl	8	0	6	1	1	0	0
Small Town or Village - Mixed	16	0	15	1	0	0	0
Total	152	0	116	29	7	0	0

Note: No schools with student participation rate below 50% were found.

Table B.36: Allocation of teacher sample in Republic of Korea

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Big City - Boy	8	0	6	2	0	0	0
Big City - Girl	8	0	6	1	1	0	0
Big City - Mixed	40	0	25	13	2	0	0
Medium or Small City - Boy	8	0	5	3	0	0	0
Medium or Small City - Girl	8	0	7	0	1	0	0
Medium or Small City - Mixed	48	0	38	6	2	2	0
Small Town or Village - Boy	8	0	6	2	0	0	0
Small Town or Village - Girl	8	0	6	1	1	0	0
Small Town or Village - Mixed	16	0	15	1	0	0	0
Total	152	0	114	29	7	2	0

Note: No schools with teacher participation rate below 50% were found.

B.19 Kosovo

- School-level exclusions consisted of schools for children with special educational needs and minority schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by region (Prishtina, Mitrovica, Peja, Prizren, Ferizaj, Gjilan,



Gjakova), resulting in seven explicit strata.

- Implicit stratification was applied by urbanization, giving a total of two implicit strata.
- Small schools were selected with equal probabilities.

Table B.37: Allocation of student sample in Kosovo

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Prishtina	42	0	42	0	0	0	0
Mitrovica	18	0	18	0	0	0	0
Peja	18	0	17	0	0	1	0
Prizren	20	0	20	0	0	0	0
Ferizaj	24	0	24	0	0	0	0
Gjilan	14	0	14	0	0	0	0
Gjakova	18	0	18	0	0	0	0
Total	154	0	153	0	0	1	0

Note: No schools with student participation rate below 50% were found.

Table B.38: Allocation of teacher sample in Kosovo

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Prishtina	42	0	35	0	0	7	0
Mitrovica	18	0	15	0	0	3	0
Peja	18	0	12	0	0	6	0
Prizren	20	0	15	0	0	5	0
Ferizaj	24	0	14	0	0	10	0
Gjilan	14	0	7	0	0	7	0
Gjakova	18	0	12	0	0	6	0
Total	154	0	110	0	0	44	0

Note: Forty-three schools were regarded non-participating as the within-school participation rate was below 50%.

B.20 Latvia

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than six students in the target grade), distance learning schools, language of instruction differs from Latvian or Russian schools and a special regime school. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual



disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.

- Explicit stratification was performed by urbanization (Riga, city, town and rural), resulting in three explicit strata.
- Implicit stratification was applied by school type, giving a total of three implicit strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies. In schools selected for both studies, different classes were to be selected for both studies.

Table B.39: Allocation of student sample in Latvia

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Riga	49	0	40	3	0	5	1
City	32	0	30	1	0	1	0
Town and Rural	70	0	60	9	0	1	0
Total	151	0	130	13	0	7	1

Note: One school was regarded non-participating as the within-school participation rate was below 50%.

Table B.40: Allocation of teacher sample in Latvia

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Riga	49	0	36	3	0	9	1
City	32	0	28	1	0	3	0
Town and Rural	70	0	56	9	0	5	0
Total	151	0	120	13	0	17	1

Note: Eleven schools were regarded non-participating as the within-school participation rate was below 50%.

B.21 Luxembourg

- School-level exclusions consisted of classes for students with intellectual disabilities and classes for immigrant students for language instructions. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school administration type (Public, Private National, Private International), resulting in three explicit strata.

- No implicit stratification.
- Planned as a census of schools and students, and a sample of teachers. One school offers two different types of curricula, and its students were therefore listed within two distinct virtual schools. All variance estimations were computed using classes as variance strata.
- Due to technical issues during the data collection, some classes and students were set to non-sampled and considered in the respective base weights. This effectively results in three schools not participating.
- Class grouping with six levels according to the ability of the students.

Table B.41: Allocation of student sample in Luxembourg

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Public	34	0	34	0	0	0	0
Privat National	5	0	4	0	0	1	0
Privat International	8	1	3	0	0	4	0
Total	47	1	41	0	0	5	0

Note: Three schools here listed as non-participating were regarded as not sampled, because of technical issues during the data collection. Two additional schools refused to participate. Considering the respective class groups in these five schools, class base weights and class adjustment factors of participating schools take these schools into account.

Table B.42: Allocation of teacher sample in Luxembourg

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Public	34	0	15	0	0	19	0
Privat National	5	0	3	0	0	2	0
Privat International	8	1	2	0	0	5	0
Total	47	1	20	0	0	26	0

Note: Twenty-one schools were regarded non-participating as the within-school participation rate was below 50%.

B.22 Malta

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than 10 students in the target grade), and language schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by sector (state, church, independent), resulting in three explicit strata.



- No implicit stratification was applied.
- Census of schools and students. All variance estimations were computed using classes as variance strata.

Table B.43: Allocation of student sample in Malta

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	
			Ineligible Schools	Sampled Schools	First Replacements	Second Replacements	Excluded Schools
Public	34	0	34	0	0	0	0
Privat National	5	0	4	0	0	1	0
Privat International	8	1	3	0	0	4	0
Total	47	1	41	0	0	5	0

Note: No schools with student participation rate below 50% were found.

Table B.44: Allocation of teacher sample in Malta

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	
			Ineligible Schools	Sampled Schools	First Replacements	Second Replacements	Excluded Schools
Church	21	0	17	0	0	4	0
Independent	7	0	4	0	0	3	0
State	14	0	10	0	0	4	0
Total	42	0	31	0	0	11	0

Note: Eleven schools were regarded non-participating as the within-school participation rate was below 50%.

B.23 The Netherlands

- School-level exclusions consisted of schools for children with special educational needs, private schools, international schools, very small schools (fewer than five students in the target grade), and schools for refugees. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by educational tracks (Vorbereidend Middelbaar Beroepsonderwijs/Praktijkonderwijs, Hoger Algemeen Voortgezet Onderwijs/ Voorbereidend Wetenschappelijk Onderwijs, mixed), resulting in three explicit strata.
- No implicit stratification was applied.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and National Assessment: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.



Table B.45: Allocation of student sample in the Netherlands

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
VMBO/PRO	54	0	5	6	3	40	0
HAVOVWO	25	0	3	1	0	21	0
Mixed	76	0	14	12	2	48	0
Total	155	0	22	19	5	109	0

Note: Three school was regarded non-participating as the within-school participation rate was below 50%.

Table B.46: Allocation of teacher sample in the Netherlands

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
VMBO/PRO	54	0	3	5	1	45	0
HAVO/VWO	25	0	2	0	0	23	0
Mixed	76	0	9	8	2	57	0
Total	155	0	14	13	3	125	0

Note: Seventeen schools were regarded non-participating as the within-school participation rate was below 50%.

B.24 Norway

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade), international schools, Samii schools, and immigrant schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by region (Oslo, other), resulting in two explicit strata.
- Implicit stratification was applied by language (two levels) and the national examination school score (four levels), giving a total of six implicit strata.
- The sample was disproportionately allocated to explicit strata.
- Small schools were selected with equal probabilities.



Table B.47: Allocation of student sample in Norway

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Oslo	40	0	37	0	0	3	0
Other	120	0	115	1	0	4	0
Total	160	0	152	1	0	7	0

Note: No schools with student participation rate below 50% were found.

Table B.48: Allocation of teacher sample in Norway

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Oslo	40	0	31	1	0	8	0
Other	120	0	107	1	0	12	0
Total	160	0	138	2	0	20	0

Note: Seventeen schools were regarded non-participating as the within-school participation rate was below 50%.

B.25 Oman

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than six students in the target grade), and remote schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by governorate (Muscat, Ash Sharqiyah North, Ash Sharqiyah South, Ad Dakhliyah, Ad Dhahirah, Al Batinah North, Al Batinah South, Al Buraimi, Musandam, Dhofar, Al Wusta) and school administration type (private, international), resulting in 13 explicit strata.
- Implicit stratification was applied by school type, giving a total of three implicit strata.
- The sample was disproportionately allocated to explicit strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.



Table B.49: Allocation of student sample in Oman

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Muscat Governorate	36	0	36	0	0	0	0
Ash Sharqiyah North Governorate	14	0	14	0	0	0	0
Ash Sharqiyah South Governorate	16	0	16	0	0	0	0
Ad Dakhliyah Governorate	26	0	26	0	0	0	0
Ad Dhahirah Governorate	10	0	10	0	0	0	0
Al Batinah North Governorate	36	0	36	0	0	0	0
Al Batinah South Governorate	24	0	24	0	0	0	0
Al Buraimi Governorate	8	0	8	0	0	0	0
Musandam Governorate	8	0	8	0	0	0	0
Dhofar Governorate	14	0	14	0	0	0	0
Al Wusta Governorate	8	0	8	0	0	0	0
Private Schools	11	0	11	0	0	0	0
International Schools	11	0	9	1	0	0	1
Total	222	0	220	1	0	0	1

Note: No schools with student participation rate below 50% were found.



Table B.50: Allocation of teacher sample in Oman

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Muscat Governorate	36	0	36	0	0	0	0
Ash Sharqiyah North Governorate	14	0	13	0	0	1	0
Ash Sharqiyah South Governorate	16	0	16	0	0	0	0
Ad Dakhliyah Governorate	26	0	26	0	0	0	0
Ad Dhahirah Governorate	10	0	10	0	0	0	0
Al Batinah North Governorate	36	0	36	0	0	0	0
Al Batinah South Governorate	24	0	23	0	0	1	0
Al Buraimi Governorate	8	0	8	0	0	0	0
Musandam Governorate	8	0	8	0	0	0	0
Dhofar Governorate	14	0	14	0	0	0	0
Al Wusta Governorate	8	0	8	0	0	0	0
Private Schools	11	0	10	0	0	1	0
International Schools	11	0	9	1	0	0	1
Total	222	0	217	1	0	3	1

Note: Three schools were regarded non-participating as the within-school participation rate was below 50%.

B.26 Portugal

- School-level exclusions consisted of very small schools (fewer than five students in the target grade) and minority-language schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by grade, school administration type (public, private) and region (Alentejo, Algarve, Lisboa, Centro, Norte, Madeira, and Açores), resulting in nine explicit strata.
- Implicit stratification was applied by type of school administration (public, private) and region (five regions), giving a total of nine implicit strata.
- Small schools were selected with equal probabilities.
- The main survey samples for ICILS 2023 and TIMSS 2023 were selected simultaneously, resulting in maximum overlap of the schools sampled for both studies. However, different classes were randomly selected for both studies.



Table B.51: Allocation of student sample in Portugal

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools		Non-participating Schools	Excluded Schools	
			Sampled Schools	First Replacements	Second Replacements		
Grade 8 only - Madeira and Açores	8	0	7	1	0	0	0
Grade 8 only – remaining regions	16	0	14	2	0	0	0
Public - Alentejo	10	0	8	2	0	0	0
Public - Algarve	10	0	9	1	0	0	0
Public - Lisboa	36	0	34	2	0	0	0
Public - Centro	28	0	26	2	0	0	0
Public - Norte	46	0	43	3	0	0	0
Private - All Regions	10	0	4	6	0	0	0
Total	164	0	145	19	0	0	0

Note: No schools with student participation rate below 50% were found.

Table B.52: Allocation of teacher sample in Portugal

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools		Non-participating Schools	Excluded Schools	
			Sampled Schools	First Replacements	Second Replacements		
Grade 8 only - Madeira and Açores	8	0	7	1	0	0	0
Grade 8 only – remaining regions	16	0	14	2	0	0	0
Public - Alentejo	10	0	8	2	0	0	0
Public - Algarve	10	0	9	1	0	0	0
Public - Lisboa	36	0	34	2	0	0	0
Public - Centro	28	0	26	2	0	0	0
Public - Norte	46	0	43	3	0	0	0
Private - All Regions	10	0	4	6	0	0	0
Total	164	0	145	19	0	0	0

Note: No schools with teacher participation rate below 50% were found.

B.27 Romania

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade) and schools with a radically different curriculum. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read



or speak the language(s) of the test.

- Explicit stratification was performed by region (Bucharest-Ilfov, north, center, south, west) and urbanization (rural, urban), resulting in eight explicit strata.
- Implicit stratification was applied by urbanization, giving a total of two implicit strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.

Table B.53: Allocation of student sample in Romania

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	
		Ineligible Schools	Sampled Schools	First Replacements	Second Replacements		Excluded Schools
Bucharest-Ilfov - All	14	1	9	1	0	2	1
North - Urban	22	0	17	1	1	3	0
Center - Urban	12	0	11	1	0	0	0
South - Urban	28	0	20	3	2	3	0
West - All	12	0	9	1	1	1	0
North - Rural	28	0	26	1	1	0	0
Center- Rural	10	0	6	2	0	2	0
South - Rural	30	2	22	1	0	4	1
Total	156	3	120	11	5	15	2

Note: Three schools were regarded non-participating as the within-school participation rate was below 50%.

Table B.54: Allocation of teacher sample in Romania

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	
		Ineligible Schools	Sampled Schools	First Replacements	Second Replacements		Excluded Schools
Bucharest-Ilfov - All	14	1	9	1	0	2	1
North - Urban	22	0	18	1	1	2	0
Center - Urban	12	0	10	1	0	1	0
South - Urban	28	0	20	3	2	3	0
West - All	12	0	9	1	0	2	0
North - Rural	28	0	24	1	1	2	0
Center- Rural	10	0	5	2	0	3	0
South - Rural	30	2	19	2	0	6	1
Total	156	3	114	12	4	21	2

Note: Eight schools were regarded non-participating as the within-school participation rate was below 50%.



B.28 Serbia

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than six students in the target grade), and schools teaching in a language other than Serbian. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by region (Belgrade, Vojvodina, Central Serbia), urbanization (city, other) and school hierarchy (Main School, Branch Department), resulting in seven explicit strata.
- Implicit stratification was applied by the national assessment performance group for mathematics, giving a total of four implicit strata.
- Small schools were selected with equal probabilities.

Table B.55: Allocation of student sample in Serbia

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools		Participating schools			Non-participating Schools	Excluded Schools
		Ineligible Schools	Sampled Schools	First Replacements	Second Replacements		
Belgrade - City - All schools	30	0	28	1	0	1	0
Belgrade - Other - All schools	8	0	8	0	0	0	0
Vojvodina - City - All schools	26	0	26	0	0	0	0
Vojvodina - Other - All schools	14	0	14	0	0	0	0
Central Serbia - City - All schools	52	0	52	0	0	0	0
Central Serbia - Other - Main School	18	0	18	0	0	0	0
Central Serbia - Other - Branch Department	8	1	7	0	0	0	0
Total	156	1	153	1	0	1	0

Note: One school was regarded non-participating as the within-school participation rate was below 50%.



Table B.56: Allocation of teacher sample in Serbia

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools		Second Replacements	Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements			
Belgrade - City - All schools	30	0	29	1	0	0	0
Belgrade - Other - All schools	8	0	8	0	0	0	0
Vojvodina - City - All schools	26	0	26	0	0	0	0
Vojvodina - Other - All schools	14	0	14	0	0	0	0
Central Serbia - City - All schools	52	0	52	0	0	0	0
Central Serbia - Other - Main School	18	0	18	0	0	0	0
Central Serbia - Other - Branch Department	8	1	7	0	0	0	0
Total	156	1	154	1	0	0	0

Note: No schools with teacher participation rate below 50% were found.

B.29 Slovak Republic

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than three students in the target grade) and language of instruction different than Slovak or Hungarian. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by language (Slovak, Hungarian, special schools) and national math assessment (high, medium and missing, low), resulting in five explicit strata.
- Implicit stratification was applied by school type (three types) and national math assessment (four levels), giving a total of seven implicit strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.



Table B.57: Allocation of student sample in Slovak Republic

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Slovak - Low	14	0	14	0	0	0	0
Slovak - Medium and Missing	53	0	53	0	0	0	0
Slovak - High	81	0	79	2	0	0	0
Hungarian - All	10	0	10	0	0	0	0
Special Schools – All	8	0	5	3	0	0	0
Total	166	0	161	5	0	0	0

Note: No schools with student participation rate below 50% were found.

Table B.58: Allocation of teacher sample in Slovak Republic

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Slovak - Low	14	0	14	0	0	0	0
Slovak - Medium and Missing	53	0	53	0	0	0	0
Slovak - High	81	0	77	2	0	2	0
Hungarian - All	10	0	10	0	0	0	0
Special Schools - All	8	0	5	3	0	0	0
Total	166	0	159	5	0	2	0

Note: Two schools were regarded non-participating as the within-school participation rate was below 50%.

B.30 Slovenia

- School-level exclusions consisted of schools for children with special educational needs, private schools with different pedagogy, home schooled students, and Italian schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- No explicit stratification was applied.
- No implicit stratification was applied.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.



Table B.59: Allocation of student sample in Slovenia

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	
			Sampled Schools	First Replacements	Second Replacements	Excluded Schools	Excluded Schools
All	175	0	163	5	0	7	0
Total	175	0	163	5	0	7	0

Note: One school was regarded non-participating as the within-school participation rate was below 50%.

Table B.60: Allocation of teacher sample in Slovenia

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	
			Sampled Schools	First Replacements	Second Replacements	Excluded Schools	Excluded Schools
All	175	0	160	5	0	10	0
Total	175	0	160	5	0	10	0

Note: Five schools were regarded non-participating as the within-school participation rate was below 50%.

B.31 Spain

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than four students in the target grade), geographically inaccessible schools, and international schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by region (Andalucía, Principado De Asturias, Canarias, Cantabria, Castilla Y León, Catalunya, Comunidad Valenciana, Comunidad Autonoma De Madrid, La Rioja, Ceuta, Melilla, Other Regions), resulting in 12 explicit strata.
- Implicit stratification was applied by school administration type (public, private) and autonomous communities (eight levels), giving a total of 16 implicit strata.
- The sample was disproportionately allocated to explicit strata.
- Small schools were selected with equal probabilities.



Table B.61: Allocation of student sample in Spain

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Andalusia	50	0	50	0	0	0	0
Asturias	50	0	49	0	0	1	0
Canary Islands	50	0	50	0	0	0	0
Cantabria	50	0	50	0	0	0	0
Castile and Leon	50	0	50	0	0	0	0
Catalonia	50	0	49	1	0	0	0
Valencia	50	0	49	0	0	1	0
Madrid	50	0	50	0	0	0	0
La Rioja	50	0	50	0	0	0	0
Ceuta	12	0	12	0	0	0	0
Melilla	8	0	8	0	0	0	0
Other Regions	40	0	40	0	0	0	0
Total	510	0	507	1	0	2	0

Note: Two schools were regarded non-participating as the within-school participation rate was below 50%.

Table B.62: Allocation of teacher sample in Spain

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Andalusia	50	0	47	0	0	3	0
Asturias	50	0	49	0	0	1	0
Canary Islands	50	0	49	0	0	1	0
Cantabria	50	0	49	0	0	1	0
Castile and Leon	50	0	49	0	0	1	0
Catalonia	50	0	48	1	0	1	0
Valencia	50	0	49	0	0	1	0
Madrid	50	0	50	0	0	0	0
La Rioja	50	0	49	0	0	1	0
Ceuta	12	0	12	0	0	0	0
Melilla	8	0	8	0	0	0	0
Other Regions	40	0	40	0	0	0	0
Total	510	0	499	1	0	10	0

Note: Ten schools were regarded non-participating as the within-school participation rate was below 50%.



B.32 Sweden

- School-level exclusions consisted of schools for children with special educational needs, special needs schools, very small schools (fewer than five students in the target grade), and international schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by average grades (high, medium, low, missing) and school administration (public, private), resulting in seven explicit strata.
- Implicit stratification was applied only for schools with missing average grade by school administration type, giving a total of two implicit strata.
- Small schools were selected with equal probabilities.
- The main survey samples for ICILS 2023 and TIMSS 2023 were selected simultaneously, resulting in a minimum overlap of the sampled schools.

Table B.63: Allocation of student sample in Sweden

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Low - Public	38	0	34	2	0	2	0
Low - Private	10	0	9	0	0	0	1
Medium - Public	40	0	40	0	0	0	0
Medium - Private	10	0	10	0	0	0	0
High - Public	38	2	34	1	0	1	0
High - Private	10	0	10	0	0	0	0
Missing - All	8	1	7	0	0	0	0
Total	154	3	144	3	0	3	1

Note: One school was regarded non-participating as the within-school participation rate was below 50%.



Table B.64: Allocation of teacher sample in Sweden

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replacements	Second Replacements		
Low - Public	38	0	33	2	0	3	0
Low - Private	10	0	9	0	0	0	1
Medium - Public	40	0	40	0	0	0	0
Medium - Private	10	0	10	0	0	0	0
High - Public	38	2	31	1	0	4	0
High - Private	10	0	10	0	0	0	0
Missing - All	8	1	7	0	0	0	0
Total	154	3	140	3	0	7	1

Note: Four schools were regarded non-participating as the within-school participation rate was below 50%.

B.33 United States

- No schools were excluded prior to school sampling. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school administration type (public, private), region (Northeast, Midwest, South, West) and poverty level (high, low), resulting in eight explicit strata.
- Implicit stratification was applied by urbanization (four levels), percentage of non-white students (two levels), and federal interagency panel on school funding (51 levels), giving a total of 408 implicit strata.
- Small schools were selected with equal probabilities.
- School sample overlap between ICILS 2023 and TIMSS 2023: The ICILS 2023 sample was selected with overlap control applying the Chowdhury approach to achieve minimum overlap between these studies.



Table B.65: Allocation of student sample in United States

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Public - Northeast - High	5	0	0	1	0	4	0
Public - Northeast - Low	20	0	7	0	2	11	0
Public - Midwest - High	6	0	5	0	0	1	0
Public - Midwest - Low	29	1	14	3	4	6	1
Public - South - High	17	2	9	1	0	5	0
Public - South - Low	51	0	30	2	4	15	0
Public - West - High	12	0	3	5	2	2	0
Public - West - Low	32	0	9	4	7	12	0
Private	11	0	3	1	2	5	0
Total	183	3	80	17	21	61	1

Note: Four schools were regarded non-participating as the within-school participation rate was below 50%.

Table B.66: Allocation of teacher sample in United States

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Public - Northeast - High	5	0	0	1	0	4	0
Public - Northeast - Low	21	0	7	1	2	11	0
Public - Midwest - High	6	0	4	0	0	2	0
Public - Midwest - Low	29	1	12	2	3	11	0
Public - South - High	17	2	5	1	0	9	0
Public - South - Low	52	0	24	2	3	23	0
Public - West - High	12	0	2	3	2	5	0
Public - West - Low	32	0	8	2	5	17	0
Private	11	0	3	1	2	5	0
Total	185	3	65	13	17	87	0

Note: Ten schools were regarded non-participating as the within-school participation rate was below 50%.

B.34 Uruguay

- School-level exclusions consisted of schools for children with special educational needs and rural schools. Within-school exclusions consisted of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.
- Explicit stratification was performed by school administration type (public, private), learning sys-

tem (Universidad del Trabajo del Uruguay UTU, Liceo), and region (capital city, rest of Uruguay), resulting in six explicit strata.

- No implicit stratification was applied.
- Small schools were selected with equal probabilities.

Table B.67: Allocation of student sample in Uruguay

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Public - Liceo - Capital City	28	1	23	0	1	3	0
Public - Liceo - Rest of Uruguay	64	0	62	2	0	0	0
Public - UTU - Capital City	10	0	10	0	0	0	0
Public - UTU - Rest of Uruguay	26	3	21	0	0	2	0
Private - Liceo - Capital City	16	0	13	3	0	0	0
Private - Liceo - Rest of Uruguay	10	1	8	1	0	0	0
Total	154	5	137	6	1	5	0

Note: Three schools were regarded non-participating as the within-school participation rate was below 50%.

Table B.68: Allocation of teacher sample in Uruguay

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Public - Liceo - Capital City	28	1	18	0	1	8	0
Public - Liceo - Rest of Uruguay	64	0	58	2	0	4	0
Public - UTU - Capital City	10	0	6	0	0	4	0
Public - UTU - Rest of Uruguay	26	3	20	0	0	3	0
Private - Liceo - Capital City	16	0	13	3	0	0	0
Private - Liceo - Rest of Uruguay	10	1	7	1	0	1	0
Total	154	5	122	6	1	20	0

Note: Eighteen schools were regarded non-participating as the within-school participation rate was below 50%.

B.35 Benchmark country: North-Rhine Westphalia (Germany)

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than three students in the target grade). Within-school exclusions consisted



of students (in regular or excluded classes) with functional or intellectual disabilities or non-native language speakers who are unable to read or speak the language(s) of the test.

- Explicit stratification was performed by school type (Gymnasium, Non-Gymnasium, and schools for children with special educational needs), resulting in three explicit strata.
- Implicit stratification was applied by and socioeconomic status (migrant proportion), giving a total of three implicit strata.
- Small schools were selected with equal probabilities.

Table B.69: Allocation of student sample in North-Rhine Westphalia (Germany)

School Participation Status – Student survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Gymnasium	42	0	38	3	0	1	0
Non-Gymnasium	70	0	65	4	0	1	0
Special education needs schools	1	0	0	1	0	0	0
Total	113	0	103	8	0	2	0

Note: No schools with student participation rate below 50% were found.

Table B.70: Allocation of teacher sample in North-Rhine Westphalia (Germany)

School Participation Status – Teacher survey							
Explicit Strata	Total Sample Schools	Ineligible Schools	Participating schools			Non-participating Schools	Excluded Schools
			Sampled Schools	First Replace-ments	Second Replace-ments		
Gymnasium	42	0	37	3	0	2	0
Non-Gymnasium	70	0	62	3	0	5	0
Special education needs schools	1	0	0	1	0	0	0
Total	113	0	99	7	0	7	0

Note: Five schools were regarded non-participating as the within-school participation rate was below 50%.

Appendix C:

CIL release module scoring rules

Table C.1 and Table C.2 provide summary information about the human and machine scoring rules used to score students responses for each task in the CIL release modules Breathing and School trip.

C.1 Breathing

Table C.1: Scoring information for CIL module Breathing

Task number	Scoring variable	Scoring rules
1	C1H01ZA	Code 1: The target file is opened by any of the methods listed below. - Method 1 (Double click): Double clicks on the file "Assignment.prs." - Method 2 (Context menu): Right clicks on the file "Assignment.prs" and then selects the menu item "[Open file]." - Method 3 (Toolbar menu): Left clicks on the file "Assignment.prs" then selects the toolbar item "[File]" and menu item "[Open]." Code 0: The target file is not opened or a non-target file is opened.
2	C1H02ZA	Code 2: The presentation file is saved as one of the names (exact match) listed below. - Breathing - Breathing.prs - 'Breathing' - 'Breathing.prs' Code 1: The presentation file is saved as any file name different to the default file name ("Assignment.prs"). Code 0: The presentation file is not saved, or the file name is not changed.
3	C1H03ZA	Code 1: Left clicks on the [WebSearch] button in the taskbar. Code 0: Other responses.
4	C1H04AC	Code 2: Gives a search phrase that (i) implies or states an explanation of the process of breathing and (ii) refers to an explanation suitable for children. E.g., How breathing works (i) explained for children (ii). Code 1: Gives a search phrase that implies or states an explanation of the process of breathing. E.g., How lungs work.

Continued on next page



Table C.1: Scoring information for CIL module Breathing (Continued)

5	C1H05ZC	Code 1: Response reflects 'No' and includes a reason from one of the categories listed below. - Category 1: The information could be exaggerated/made up to sell a product. - Category 2: Only one anonymous customer testimonial. - Category 3: Lack of independent research. - Category 4: Lack of cited sources or supporting evidence. Code 0: Response reflects 'Yes' (with any explanation). OR States that the site serves as promotional material without explicitly referring to the pills, a 'product,' or equivalent. OR Response is vague, irrelevant, or repeats the stem.
6	C1H06ZC	Code 1: Response includes a reason from one of the categories listed below. - Category 1: (Yes) The author's expertise (university professor). - Category 2: (Yes) The information can be edited by the public (public scrutiny). - Category 3: (No) The information might be false or fake because it can be edited by the public. - Category 4: (No) There is no verifiable evidence of the author's identity or expertise. Code 0: Includes reasons from any of the categories listed below. OR Response is vague, irrelevant, or repeats the stem. - Category 1: (Yes) The scientific nature of the information. - Category 2: (Yes) The source of the information is known (without referring to the author's expertise).
7	C1H07AL	Presentation title - Code 1: Content of title is appropriate: Title refers to the concept of "Breathing." - Code 0: No title or content of title is inappropriate OR other titles that fail to include the concept of "Breathing."

Continued on next page



Table C.1: Scoring information for CIL module Breathing (Continued)

7	C1H07BL	Text – Information design (distinction between headings and text body) Code 2: Clear control of layout There is clear evidence of control and planning in the layout of the presentation. Formatting tools have been used successfully to support readers' understanding of the role of the different text elements on all slides that use text (excluding the title slide). - There is a clear distinction between main body text and heading using bolding, underlining, italicizing or increased font size for the heading(s) OR placing the heading/sub-headings on a separate slide. Code 1: Basic control of layout - There is some evidence of planning in the layout of the presentation. Formatting tools have been used to some degree to support readers' understanding of the role of the different text elements: Note: Code 1 may include inconsistent use of formatting. It may include clear control of some elements with some basic or poor control of others. - Headings and sub-headings are mostly distinguishable from the main body text by formatting, e.g., using color, bolding, underlining, italicizing, or increased font size OR placing the heading/sub-headings on a separate slide. OR - ONLY (paragraph) spacing has been used to distinguish between headings and body text. OR Text and headings have been copied and pasted from the web resources, maintaining the formatting that distinguishes between headings and body text from the original resource. Code 0: Little or no control of layout There is little or no evidence of control and planning in the layout of the text. Formatting tools do not support readers' understanding of the role of the different text elements: - The text layout does not distinguish between headings, sub-headings, and main body text.

Continued on next page



Table C.1: Scoring information for CIL module Breathing (Continued)

7	C1H07CL	Text position (alignment)
		Code 2: Clear control of layout
		There is clear evidence of control and planning in the layout of the presentation.
		- The position and alignment of text makes it easy to follow the information visually (such as spacing between paragraphs or use of bullet points to break up paragraphs).
		Code 1: Basic control of layout
		There is some evidence of planning in the layout of the presentation.
		Note: Code 1 may include inconsistent use of formatting. It may include clear control of some elements with some basic or poor control of others.
		- The text is mostly positioned thoughtfully and mostly well aligned. Some text may be close together but does not overlap with other text or the images or inhibit reading flow.
		Code 0: Little or no control of layout
		There is little or no evidence of control and planning in the layout of the text.
		- Text is not well aligned or may overlap (with other text or with the images), inhibiting reading flow.
7	C1H07DL	Images – Layout
		Code 2: Clear control of image layout
		The images are well aligned and appropriate in size.
		Code 1: Basic control of image layout
		There is some evidence of the layout of the image(s). One image may be too big, too small, or overlap with other images.
		Code 0: Little or no control of image layout
		There is little or no evidence of planning the layout of the images. The image(s) are too big, too small, or overlap with other images.

Continued on next page

Table C.1: Scoring information for CIL module Breathing (Continued)

7	C1H07EL	Text – Contrast
		Code 2: Clear and consistent contrast Slides show evidence of careful planning regarding the use of color to support contrast between the slide contents and backgrounds. - There is sufficient contrast to enable all text to be seen and read easily. Code 1: Inconsistent contrast Slides show some evidence of planning regarding the use of color to support contrast between the slide contents and backgrounds. - The text mostly contrasts sufficiently with the background to be seen and read easily. Code 0: Little or no contrast Slides show no evidence of planning regarding the use of color to support contrast between the slide contents and backgrounds. - The grey background and text colors have been used throughout. OR - The colors of the headings, text and backgrounds show poor contrast throughout the presentation and inhibit reading and viewing of text.

Continued on next page



Table C.1: Scoring information for CIL module Breathing (Continued)

7	C1H07FL	Color – Continuity Code 2: Careful continuity planning Slides show evidence of careful planning regarding the use of color to demonstrate the relationships of elements within and across the slides. - Colors of headings, body text, or backgrounds across the slides are consistent. OR - Colors of headings, body text, or backgrounds across the slides change with clear reason (e.g., to emphasize different processes). Code 1: Some continuity planning Slides show some evidence of planning regarding the use of color to demonstrate the relationships of elements within and across the slides. - There is some use of color to denote specific purposes of text (such as to denote headings and text body). This may be evident within but not across slides, or across slides but with minor inconsistencies. Code 0: Little or no continuity planning Slides show no evidence of planning regarding the use of color or the use of color is jarring/distracting. - The default grey background and text color scheme has been used. OR - The colors of the headings, text and backgrounds seem to have been assigned randomly or decoratively without any clear connection to the purpose of the text within or across the slides. OR - The background color is different for each slide with no other use of color (this is deemed as insufficient evidence of planning). OR - The presentation consists of a single slide with content (blank slides—without content—should be ignored).

Continued on next page



Table C.1: Scoring information for CIL module Breathing (Continued)

7	C1H07GL	Images – Relevance
		Code 2: At least two steps are clearly supported by relevant images only. Two or three steps of breathing are introduced by headings or explained with text and each step is supported by the inclusion of a relevant image only. No irrelevant images have been included. Code 1: One step and no irrelevant images or more than one step and any irrelevant images. One of the three steps of breathing is introduced by a heading or explained with text and supported by a relevant image and no irrelevant images have been included. OR Two or three steps of breathing are introduced by headings or explained with text and each step is supported by the inclusion of a relevant image AND any irrelevant images have been included. Code 0: One step and any irrelevant images or no steps are clearly explained by images. One of the three steps of breathing is introduced by a heading or explained with text and supported by a relevant image AND any irrelevant images have been included. OR Images included are either irrelevant or not linked to any of the stages of breathing. OR No images have been included.

Continued on next page



Table C.1: Scoring information for CIL module Breathing (Continued)

7	C1H07HL	Ideas – Completeness and sequencing
		Code 2: Includes all three steps of breathing in a coherent sequence.
		1) Breathing in
		2) Gas exchange
		3) Breathing out
		OR
		1) Breathing out
		2) Breathing in
		3) Gas exchange
		OR
		1) Breathing in
		2) Breathing out
		3) Gas exchange
		Code 1: Includes two steps of breathing in a coherent sequence.
		1) Breathing in
		2) Gas exchange
		OR
		1) Gas exchange
		2) Breathing out
		OR
		1) Breathing in
		2) Breathing out
		OR
		1) Breathing out
		2) Breathing in
		Code 0: Includes fewer than two steps of breathing or includes more than one step in an incoherent sequence.
		1) Gas exchange
		2) Breathing in
		3) Breathing out
		OR
		1) Gas exchange
		2) Breathing out
		3) Breathing in
		OR
		1) Breathing out
		2) Gas exchange
		3) Breathing in
		OR
		1) Breathing out
		2) Gas exchange
		OR
		1) Gas exchange
		2) Breathing in
		OR
		1) Gas exchange
		OR
		1) Breathing in
		OR
		1) Breathing out

Continued on next page



Table C.1: Scoring information for CIL module Breathing (Continued)

7	C1H07IL	Information – Accuracy
		Code 1: Information provided is accurate The information given about each step in the breathing process that is presented is accurate. See the marking resource below for what constitutes an accurate description for a step. Code 0: Information provided contains inaccuracies There are errors or omissions in the descriptions of the steps that are presented about breathing.
7	C1H07JL	Information – Adaptation
		Code 1: Some information adapted The presentation conveys the steps of breathing in a way that is partly accessible to an 8–9 year old audience: - Some scientific terms have been adapted or defined in plain language. - There is little explicit detail of any biological and chemical processes. Code 0: Information not adapted Makes no attempt to address the needs of an 8–9 year old audience: - No scientific terms have been adapted. - Biological or chemical processes may have been included without explanation in plain language.



C.2 School trip

Table C.2: Scoring information for CIL module School trip

Task number	Scoring variable	Scoring rules
1	C1S01ZA	Code 1: Navigates to the [m-town] tours website by any of the following methods: - Left clicks the target hyperlink - Right clicks the target hyperlink and selects one of: - Open - Open Link in New Window - Open Link in New Tab - Types or copies and pastes the URL from the target link into the browser's address bar (current browser or in new tab / window) Note: The target hyperlink is www.[m-town]-tours.icils/walking_tour3/places.
2	C1S02ZA	Code 2: Data is appropriately filtered by two criteria (by price as [\$6-10] and by Maximum group size as 21 to 30) and clicks on the hyperlink for Heritage Museum. Code 1: Data is appropriately filtered by one criterion (by price as [\$6-10] or by Maximum group size as 21 to 30) and clicks on the hyperlink for Heritage Museum. Code 0: Other responses
3	C1S03ZA	Code 1: Clicks on the website navigation tab labeled "Spreadsheet." Code 0: Other responses
4	C1S04AA	Spreadsheet value Code 1: Enters [\$5] or [5] into any spreadsheet cell. Code 0: Other responses
4	C1S04BA	Spreadsheet cell Code 1: Enters anything into cell B7. Code 0: Other responses
5	C1S05ZM	Code 1: Selects option B ("=[Sum](C5:C8)"). Code 0: Other responses
6	C1S06ZMA	Code 1: Chooses "Yes." Code 0: Chooses "No" or other responses.
6	C1S06ZMB	Code 1: Chooses "No." Code 0: Chooses "Yes." or other responses.

Continued on next page



Table C.2: Scoring information for CIL module School trip (Continued)

6	C1S06ZMC	Code 1: Chooses "Yes." Code 0: Chooses "No." or other responses.
7	C1S07ZA	Code 1: Clicks on one of the hyperlinks 'Free [Itinerary Builder]' or 'www.[itinerary-builder].icils' Code 0: Other responses
8	C1S08AL	Appropriate title Code 1: The content of the title includes 'Walking Tour' or 'Trip' or 'Sightseeing' or implies an excursion, walking tour or sightseeing activity. Code 0: The content of title is irrelevant, insufficient or inappropriate or no title is included.
8	C1S08BL	Relevant start and finish time Code 1: Relevant times are included Includes the start time ([9 am]) and finish time ([3 pm]) of Walking Tour 3. Code 0: Relevant times are not included Includes incorrect start and/or finish times or fails to include the start and/or finish time.

Continued on next page



Table C.2: Scoring information for CIL module School trip (Continued)

8	C1S08CL	Relevant costs Code 2: Total cost is included The information sheet includes any one of the total costs of: - The whole trip ([52]). - Walking Tour 3 and entry to the Heritage Museum ([11]). - Walking Tour 3 and entry to the Heritage Museum and lunch ([23]). - Walking Tour 3 and entry to the Heritage Museum and Train Ticket ([40]). Code 1: Relevant costs are included (without total) Includes relevant costs of both Walking Tour 3 ([5]) and entry to the Heritage Museum ([6]). Note: May also include the cost of lunch ([12]) and/or the cost of the train ticket ([29]). Code 0: Relevant costs are not included Includes incorrect costs for one or both activities or fails to include one or both costs.
8	C1S08DL	Map route – Accuracy Code 3: Map route is accurate The map shows the direct route from [Pier 3] along [Hoffer St] to [Miller Rd], past the Museum and back to [Pier 3] via [East St], [Edmont St] and [Southern Promenade]. Code 2: Map route goes from [Pier 3] to the Museum (along specified route) and back to [Pier 3] (by any route that can be walked) The map shows the route from [Pier 3] along [Hoffer St] to [Miller Rd], at least up to the Museum, and back to [Pier 3] via any route that could be walked. The map may show the direct route (see Code 3) with only [Edmont St] omitted. There are no extra paths drawn. Code 1: Map route goes from [Pier 3] to Museum only (by any route that can be walked) The map shows any route that can be walked from [Pier 3] to the Museum but does not return to [Pier 3]. Code 0: Map route does not go to the Museum or is nonsensical

Continued on next page



Table C.2: Scoring information for CIL module School trip (Continued)

8	C1S08EL	Information – Location and relevance
		Code 2: Includes only the two new exhibitions
		Refers to the exhibitions “History of [M-town]” and “[M-town] Musicians.” May rephrase or copy and paste information from website.
		Code 1: Includes the two new exhibitions and either or both of the two old exhibitions or includes one new exhibition (with or without either or both of the old exhibitions)
		Refers to “History of [M-town],” “[M-town] Musicians” and either or both of “Film and TV” and “Eco Fashion in [M-town].”
		OR
		Refers to one of “History of [M-town]” or “[M-town] Musicians.” May rephrase or copy and paste information from website.
		Code 0: Does not include either of the two new exhibitions
		Fails to include either of the exhibitions “History of [M-town]” or “[M-town] Musicians.”

Continued on next page



Table C.2: Scoring information for CIL module School trip (Continued)

8	C1S08FL	Information sheet – Layout
		Code 2: Clear control of layout
		There is clear evidence of control and planning in the layout of the information sheet. Formatting tools have been used successfully to support readers' understanding of the role of the different text elements.
		There is a clear distinction between main body text and sub-headings using bolding, underlining, italicizing or increased font size for the heading(s).
		The position and alignment of text boxes make it easy to follow the information visually (such as spacing between paragraphs or use of bullet points to break up paragraphs).
		Code 1: Basic control of layout
		There is some evidence of planning in the layout of the information sheet. Formatting tools have been used to some degree to support readers' understanding of the role of the different text elements.
		Example characteristics of code 1 include:
		- Headings and sub-headings are mostly distinguishable from the main body text by formatting, e.g., using color, bolding, underlining, italicizing or increased font size. - The text boxes are mostly positioned thoughtfully and are mostly well aligned. Some text boxes may be close together but do not overlap with each other or overlap with the map or inhibit reading flow. - All text that is not the title uses the same formatting. - Fewer than two of the necessary information elements (time, cost, exhibitions) have sub-headings.
		Code 0: Little or no control of layout
		There is little or no evidence of control and planning in the layout of the text. Formatting tools do not support readers' understanding of the role of the different text elements.
		The text layout does not distinguish between headings, sub-headings, and main body text.
		OR
		Text boxes are not well aligned or may overlap (with each other or with the map), inhibiting reading flow.
		OR
		Only a title and map are present (so there is not enough evidence to assess the work).

Continued on next page



Table C.2: Scoring information for CIL module School trip (Continued)

8	C1S08GL	Information sheet – Text contrast
		Code 2: All text and titles are clear and legible There is sufficient contrast between the background color and text to support readability for all text including the title. Code 1: Some of the text and titles are clear and legible There is sufficient contrast between the background color and text to support readability for some but not all the text and/or title. Code 0: Difficult to read The contrast between the background color and the text and title is poor. This inhibits the readability of the text AND the title. OR No text or headings/titles have been included.

Appendix D:

CT release module scoring rules

Table D.1 and Table D.5 provide summary information about the human and machine scoring rules used to score students responses for each task in the CT release modules Automated bus and Farm drone.

This appendix includes an explanation of the conceptual basis and operationalization of the scoring of students' block-based code algorithms in the Farm drone module used in ICILS 2018 and ICILS 2023.

D.1 Automated bus

Table D.1: Scoring information for CIL module Automated bus

Task number	Scoring variable	Scoring rules
1	T2A01ZA	Codes 01 to 07 refer to 1 to 7 correctly sequenced rows (before the first error).
2	T2A02AA	Route nodes Codes 01 to 011 refer to 1 to 11 unique nodes visited (excluding the Start node).
2	T2A02BA	Links between route nodes Codes 000 to 898 refer to the number of links between nodes (as shown on the counter).
2	T2A02CA	Objective Code 2: Final selected node is "Sports Event." Code 1: Final selected node is adjacent to the "Sports Event" node ("Female2" or "Female4"). Code 0: Other responses

Continued on next page



Table D.1: Scoring information for CIL module Automated bus (Continued)

2	T2A02ZA	Composite of route nodes and links Code 3: Route involves exactly 11 unique nodes, exactly 11 links, and concludes at the "Sports Event" node (T2A02AA=11 T2A02BA=11 T2A02CA=2). Code 2: Route includes between 7 and 11 unique nodes, and the number of links does not exceed the number of unique nodes plus three. The route must either end at a "Sports Event" node or finish on an adjacent node such as "[Female2]" or "[Female4]" (T2A02AA=7 to 11 & T2A02BA ≤ T2A02AA+3 & T2A02CA=1 or T2A02CA=2). Code 1: Route includes between 6 and 11 unique nodes (excluding the Start node) and has a total of 17 or fewer links connecting these nodes (T2A02AA=6 to 11 T2A02BA = 17). Code 0: Other responses
3	T2A03ZA	Code 2: The attempt number given by the student corresponds with a row where the value for Total Time is 30 mins. Code 1: The attempt number provided by the student corresponds to a row where the Total Time value is the lowest among all rows, excluding instances where the time is exactly 30 minutes. Code 0: The attempt number given by the student corresponds with a row where the Total Time is not the lowest value compared to other rows or a result in the table is present, but no attempt number is given.
4	T2A04ZA	Code 1: Diagram 1 Code 0: Other responses
5	T3A05ZA	Code 3: Syntax and logic correct Code 2: Syntax correct, logic incorrect Code 1: Logic correct, syntax incorrect Note: Logic refers to the hierarchical flow of information. Syntax refers to the specific commands (Yes/No, Stop at next bus stop) being used in the correct part of the decision tree.

Continued on next page



Table D.1: Scoring information for CIL module Automated bus (Continued)

6	T3A06AA	Simulation – decision tree
		Code 4: Decision tree configuration reflects the conditions nighttime and raining.
		Code 3: Decision tree configuration reflects the conditions nighttime and not raining.
		Code 2: Decision tree configuration reflects the conditions not nighttime and raining.
		Code 1: Decision tree configuration reflects the conditions not nighttime and not raining.
		Code 0: Other responses
6	T3A06BA	Simulation – distance
		Codes 01 to 11 correspond to the options 0[m] to 1000[m].
7	T3A07ZC	Code 2: The method described in the response demonstrates a systematic approach of repeated trials that helped the student converge on the correct distance.
		Code 1: Response describes a process that does not indicate systematic trialing or calculation and is unsystematic/random with a stop condition (i.e., refers to the end of the process where the result was determined).
		Code 0: Response indicates that no strategy was used or is vague, irrelevant, or repeats the stem.

Continued on next page



Table D.1: Scoring information for CIL module Automated bus (Continued)

8	T3A08ZC	Code 2: Response includes content from two different categories listed below. - Category 1: Real world systems can be dangerous to engage/simulations can be safer to run than real world trials. - Category 2: Some real-world systems cannot be engaged/cannot be easily created or observed. - Category 3: Simulations allow for easier testing/controlling of variables. - Category 4: Simulations allow for large amounts of data to be collected efficiently/quickly. - Category 5: Simulations cost less money to run than engaging real world systems Code 1: Response includes content from one of the categories listed above. Note: This includes responses in which different explanations from the same category are given. Code 0: Response is vague, irrelevant or repeats the stem.



D.2 Farm drone

Algorithm construction tasks

In the algorithm construction tasks, students were required to select and sequence block-based commands for the farm drone to perform specified actions. These actions included:

- Moving the drone to specified tile locations.
- Dropping water, seeds, or fertilizer on dirt, small crop, and large crop targets.
- Applying conditional logic based on crop size (using an “if-do” command).

After two warm-up tasks, students were introduced to the “repeat” command, enabling multiple repetitions of command sequences.

Algorithm debugging tasks

The algorithm debugging tasks required students to modify a pre-configured sequence of commands to correct the drone’s behavior according to the specified task objective(s). These tasks typically required one or two minor changes to the code configuration, though the work space was unconstrained and students had the option to remove all commands and create an entirely new algorithm.

Task complexity

As students progressed through the module, task complexity increased based on the number of actions needed to be completed by the drone and the number, type, and arrangement of targets (where water, seeds, or fertilizer needed to be dropped).

Criteria for scoring

Students’ solutions for both algorithm construction and debugging tasks were evaluated based on two main criteria: correctness and efficiency. The ICILS research team identified variables and developed specific scoring logic for each task. The measures for correctness and efficiency were then combined into a single score for each coding task, which was generated automatically by the test delivery system.

Below, we outline the scoring logic for both criteria and explain how they were combined to form the overall score for each task.

Scoring correctness

Correctness was determined by how well the drone’s actions aligned with the task instructions. In simpler tasks, correctness was evaluated based on whether the drone reached a specified location. In more complex tasks that involved both movement and target actions, correctness was evaluated based on three sub-criteria, which were composed of two variables: correct targets and incorrect targets:

1. Correct targets: The number of targets with the correct materials (2 score points per target).
2. Correct targets with incorrect materials: The number of targets with incorrect materials (1 score point per target).
3. Incorrect targets: Whether any materials were dropped on non-target tiles (1 score point if no materials were dropped on incorrect targets; 0 score points if any materials were dropped on incorrect targets).

Each solution was first scored on these variables (correct targets and incorrect targets). Then, the frequency of the scores and their combinations for all solutions were analyzed iteratively to establish a scoring logic for correctness on each task. This logic combined the correct and incorrect target scores into a single initial combined correctness score.

After the scoring logic was established for each task, a Rasch item response theory (Rasch, 1960) scaling



analysis (see [Appendix 11](#) for further details) was performed to assess the measurement validity of the scoring logic. Where necessary, we recoded scoring categories for certain items.

[Table D.2](#) shows an example of how this scoring logic was applied, along with the recoding of categories following initial scaling.

For each task, the score for a completely correct response remained as the highest unique score category. The recoding of score categories for partially correct responses varied across tasks and was informed by response frequencies, differences in response quality, and the extent to which these differences were reflected in the student data.

Table D.2: Example scoring of the correctness of student coding responses

Response description (combining targets and incorrect targets)	Targets score	Incorrect targets score		Initial combined score		Recoded combined score
All 4 targets reached with only correct materials and no incorrect targets	8	1	→	4	→	2
All 4 targets reached with only correct materials and one or more incorrect targets	8	0	→	3	→	1
At least 2 targets with only the correct materials OR All 4 targets with the incorrect material and no incorrect targets	4, 5, 6, or 7	1	→	2	→	1
At least 2 targets with only the correct materials OR All 4 targets with the incorrect material and one or more incorrect targets	4, 5, 6, or 7	0	→	1	→	0
Fewer than 2 targets met	Fewer than 4	0 or 1	→	0	→	0

Scoring efficiency

Efficiency was based on the number of commands used in the solution. The most efficient solution for each task used the fewest commands. For example, if a task required the drone to move forward five times, this was most commonly achieved by students using either five “move forward” commands in sequence or using one “repeat” command (configured to 5 repeats) and one “move forward” command. The latter solution which is the most efficient, uses two commands. In theory other configurations are also possible, such as to use one “move forward” command and a “repeat” command (configured to 4 repeats) with an additional “move forward” command. This solution would use three commands.

For each task it was possible to identify clusters of the number of commands that most likely corresponded to different approaches to solving the task. For each task we created an efficiency score hierarchy based on the number of commands used in the solution and informed by the task requirements, the likely characteristics of solutions that could be associated with different numbers of commands, and the frequencies of the number of commands across the responses.

[Table D.3](#) shows an example of how this scoring logic was applied, together with the recoding of categories following initial scaling.



Table D.3: Example scoring of the efficiency of student coding responses

Comment on efficiency	Commands	Responses (%) [*]		Code
Minimum necessary commands	6	28	→	3
Repeat with additional commands	7 - 10	16	→	2
11 commands was the minimum necessary if not using the repeat	11 - 15	41	→	1
More than 15 commands showed a good deal of redundancy	15	6	→	0

Note:

^{*} These percentages include all responses using the given number of commands regardless of their correctness.

Combining correctness and efficiency

To produce a final score for each task, the efficiency score was used to adjust the correctness score. However, we decided that the efficiency of a solution should only be used to adjust the highest solution score for each task. Operationally this meant that the efficiency of code was not considered to be important in evaluating the quality of incorrect or partially correct responses, but that it was considered to be important in identifying differences in the quality of fully correct responses.

We created a new composite variable (correctness and efficiency) for each coding task. For this variable, the score for fully correct responses was adjusted such that:

- fully correct responses with the highest efficiency score were adjusted to one score point higher than the highest correctness score;
- fully correct responses with a partial credit (moderate) efficiency score were not adjusted; and
- fully correct responses with a zero efficiency score were adjusted to one score point lower than the highest correctness score.

Incorrect or partially correct responses were not adjusted based on efficiency. [Table D.4](#) shows an example of how correctness and efficiency scores were combined to form the single composite variable.



Table D.4: Example scoring of correctness and efficiency combined

Correctness score	Efficiency score	Conceptual description of correctness and efficiency	Recoded combined score (CE)
2	3	All 4 targets using only the correct material with no irrelevant targets using no more than 6 commands	3
2	2,1	All 4 targets using only the correct material with no irrelevant targets using between 7 and 15 commands	2
2	0	All 4 targets in one row reached using only the correct material with no irrelevant targets using more than 15 commands	1
1	N/A	At least 2 targets with only the correct materials and no irrelevant targets using any number of commands OR All 4 targets with the incorrect material and no irrelevant targets using any number of commands	1
0	N/A	Fewer than 2 targets met	0

Table D.5: Scoring information for CIL module Farm drone

Task number	Scoring variable	Scoring rules
1	T2F01E	Code 3: Objective was achieved with one command Code 2: Objective was achieved with two to five commands Code 1: Objective was achieved with more than five commands Code 0: Other responses
2	T2F02L	Code 2: Exact target coordinates reached Code 1: Adjacent to target coordinate reached Code 0: Other responses
3	T2F03TEC	Code 2: All four targets were reached using only correct materials with no irrelevant targets, using three commands. Code 1: All four targets were reached using only correct materials with no irrelevant targets, using no more than eight commands. Code 0: Other responses

Continued on next page



Table D.5: Scoring information for CIL module Farm drone (Continued)

4	T3F04TEC	Code 3: All four targets were reached using only the correct material with no irrelevant targets, using no more than six commands. Code 2: All four targets were reached using only the correct material with no irrelevant targets, using seven to ten commands. Code 1: Response reflects one of the outcomes listed below. - All four targets were reached using only the correct material with no irrelevant targets, using more than 15 commands. - At least two targets were reached using only the correct material with no irrelevant targets, using any number of commands. - All four targets were reached using incorrect materials with no irrelevant targets, using any number of commands. Code 0: Other responses
5	T2F05TEC	Code 3: All eight targets were reached using only the correct material with no irrelevant targets, using 12 to 14 commands. Code 2: All eight targets were reached using only the correct material with no irrelevant targets, using 15 to 22 commands. Code 1: Response reflects one of the outcomes listed below. - All eight targets were reached using only the correct material with no irrelevant targets, using 23 or more commands. - At least four targets were reached using only the correct material, but includes one or more irrelevant targets, using any number of commands. Code 0: Other responses
6	T3F06TEC	Code 3: All six targets were reached using only the correct materials with no irrelevant targets, using no more than five commands. Code 2: All six targets were reached using only correct materials with no irrelevant targets, using six to seven commands. Code 1: Response reflects one of the outcomes listed below. - All six targets were reached using only the correct materials with no irrelevant targets, using more than 13 commands. - At least three targets were reached using only the correct materials with no irrelevant targets, using any number of commands. - At least two targets were reached using the correct materials and two targets using the incorrect materials with no irrelevant targets, using any number of commands. Code 0: Other responses

Continued on next page



Table D.5: Scoring information for CIL module Farm drone (Continued)

7	T2F07TEC	Code 3: All 12 targets were reached using only the correct materials with one or more irrelevant targets, using no more than 19 commands. Code 2: All 12 targets were reached using only the correct materials with one or more irrelevant targets, using no more than 24 commands. Code 1: Response reflects one of the outcomes listed below. - At least eight targets were reached using only the correct materials with no irrelevant targets. - At least four targets were reached using only the correct materials and eight targets using the incorrect materials with no irrelevant targets. - All 12 targets were reached using only the correct materials with one or more irrelevant targets, using no more than 24 commands. - At least six targets were reached using only the correct materials with no irrelevant targets, using any number of commands. - At least two targets were reached using only the correct materials and 10 targets using the incorrect materials with no irrelevant targets, using any number of commands. Code 0: Other responses
8	T2F08TEC	Code 2: Response reflects one of the outcomes listed below. - All 16 targets were reached using only the correct materials with no irrelevant targets, using no more than 15 commands. - All 16 targets were reached using only the correct materials with no irrelevant targets, using 16 to 22 commands. Code 1: Response reflects one of the outcomes listed below. - All 16 targets were reached using only the correct materials with no irrelevant targets, using more than 42 commands. - At least seven targets were reached using only the correct materials with no irrelevant targets, using any number of commands. - At least two targets were reached using only the correct materials and 10 targets with the incorrect materials with no irrelevant targets, using any number of commands. Code 0: Other responses



References

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Appendix E:

Background variables used for conditioning

The table lists all the international student-level variables (along with their respective scoring) that were used in the conditioning of plausible values for CIL and CT. To account for between-school differences, stratum indicator, the average of weighted likelihood ability (WLE) estimates for all other students in the same school were introduced as direct conditioning variables. Students' gender was also included as a direct conditioning variable. In addition, all available international student questionnaire variables were included to condition on students' background information. These categorical variables were dummy-coded prior to principal component analysis (PCA). This step extracted a smaller set of uncorrelated components that captured most of the variance in the original variables. These components were subsequently used as direct conditioning variables in the model.



Table E.1: Background variables used for conditioning

Variable	Name	Values	Regressor
Stratum	IDSTRATE [†]		Direct
Adjusted school mean CIL achievement	SCH_MN_CIL	Continuous variable	Direct
Adjusted school mean CT achievement	SCH_MN_CT	Continuous variable	Direct
Gender	GENDER	1=Female 2=Male 9=Missing	Direct
Age	S_AGE	Continuous variable	PCA
Highest level of parental occupation	S_HISEI	Continuous variable	PCA
Highest level of parental education	S_HISCED	1=[ISCED level 6, 7, or 8] 2=[ISCED level 4 or 5] 3=[ISCED level 3] 4=[ISCED level 2] 5=I do not expect to complete [ISCED level 2] 9=Missing	PCA
Student expected levels of education to complete	S_ISCED	1=[ISCED level 6, 7, or 8] 2=[ISCED level 4 or 5] 3=[ISCED level 3] 4=[ISCED level 2] 5=I do not expect to complete [ISCED level 2] 9=Missing	PCA
Country of birth - You	IS3G04A [†]	1=[Country of test]	PCA
Country of birth - Parent or guardian 1	IS3G04B [†]	2=[Other country/Group A]	
Country of birth - Parent or guardian 2	IS3G04C [†]	3=[Other country/Group B] 4=[Another country] 9=Missing	
Language at home	IS3G05 [†]	1=[Language of test] 2=[Other language 1] 3=[Other language 2] 4=[Another language] 9=Missing	PCA

Note: [†] Length of values varies across countries.



Table E.1: Background variables used for conditioning (cont'd)

Variable	Name	Values	Regressor
Books at home	IS3G14	1=None or very few 2=Enough to fill one shelf 3=Enough to fill one bookcase 4=Enough to fill two bookcases 5=Enough to fill three or more bookcases 9=Missing	PCA
Home internet connection - Drop out	IS3G15A	1=There is no internet connection in my home	PCA
Home internet connection - Slow	IS3G15B	2=Never or almost never 3=At least once a week but not every day 4=Once a day 5=More than once a day 9=Missing	
Home devices - Desktop or laptop	IS3G16AA	1=None	
Home devices - Tablet devices	IS3G16AB	2=One 3=Two 4=Three or more 9=Missing	
Home devices for schoolwork - Desktop or laptop	IS3G16BA	1=Never	
Home devices for schoolwork - Tablet devices	IS3G16BB	2=Sometimes	PCA
Home devices for schoolwork - Smartphone	IS3G16BC	3=Most times 4=Always 9=Missing	
How long have you been using computers	IS3G17	1=Less than a year 2=At least 1 year but less than 3 years 3=At least 3 years but less than 5 years 4=At least 5 years but less than 7 years 5=7 years or more 9=Missing	PCA
ICT on school days - at school for schoolwork	IS3G18A	1=Never	PCA
ICT on school days - at school for other purposes	IS3G18B	2=Less than once a month	
ICT on school days - Outside for schoolwork	IS3G18C	3=At least once a month but not every week	
ICT on school days - Outside for other purposes	IS3G18D	4=At least once a week but not every day	
ICT on non-school days - Outside for schoolwork	IS3G18E	5=Every day, less than one hour	
ICT on non-school days - Outside for other purposes	IS3G18F	6=Every day, at least one hour but less than two hours 7=Every day, at least two hours but less than three hours 8=Every day, three hours or more 9=Missing	



Table E.1: Background variables used for conditioning (cont'd)

Variable	Name	Values	Regressor
Limit screen time on school days	IS3G19A	1=Yes - and it includes the time I spend doing schoolwork on a screen	PCA
Limit screen time on non-school days	IS3G19B	2=Yes - but it does not include the time I spend doing schoolwork on a screen	
		3=No - there is no limit	
		9=Missing	
At school - find information online	IS3G20AA1	1=To a large extent	PCA
At school - refine searches	IS3G20AB1	2=To a moderate extent	
At school - evaluate information reliability	IS3G20AC1	3=To a small extent	
At school - include references	IS3G20AD1	4=Not at all	
At school - identify scam messages	IS3G20AE1	9=Missing	
At school - manage privacy settings	IS3G20AF1		
At school - organize files	IS3G20BA1	1=To a large extent	PCA
At school - edit documents	IS3G20BB1	2=To a moderate extent	
At school - edit digital media	IS3G20BC1	3=To a small extent	
At school - use spreadsheets	IS3G20BD1	4=Not at all	
At school - create programs	IS3G20BE1	9=Missing	
At school - write programs	IS3G20BF1		
Outside school - find information online	IS3G20AA2	1=To a large extent	PCA
Outside school - refine searches	IS3G20AB2	2=To a moderate extent	
Outside school - evaluate information reliability	IS3G20AC2	3=To a small extent	
Outside school - include references	IS3G20AD2	4=Not at all	
Outside school - identify scam messages	IS3G20AE2	9=Missing	
Outside school - manage privacy settings	IS3G20AF2		
Outside school - organize files	IS3G20BA2	1=To a large extent	PCA
Outside school - edit documents	IS3G20BB2	2=To a moderate extent	
Outside school - edit digital media	IS3G20BC2	3=To a small extent	
Outside school - use spreadsheets	IS3G20BD2	4=Not at all	
Outside school - create programs	IS3G20BE2	9=Missing	
Outside school - write programs	IS3G20BF2		
Learning responsible social media use	IS3G20CA	1=To a large extent	PCA
Learning to recognize cyberbullying	IS3G20CB	2=To a moderate extent	
Learning to report cyberbullying	IS3G20CC	3=To a small extent	
Learning physical health and ICT use	IS3G20CD	4=Not at all	
Learning psychological health and ICT use	IS3G20CE	9=Missing	
Text chat with others	IS3G21A	1=Never	PCA
Use social media or view posts	IS3G21B	2=Almost never	
Check social media	IS3G21C	3=Sometimes	
Find information online	IS3G21D	4=Often	
Watch videos, streams, or TV	IS3G21E	5=Very often	
Listen to music or radio	IS3G21F	9=Missing	



Table E.1: Background variables used for conditioning (cont'd)

Variable	Name	Values	Regressor
In lessons – Test language	IS3G22A	1=I don't study this subject/these subjects	PCA
In lessons – Other national languages	IS3G22B	2=Never	
In lessons – Mathematics	IS3G22C	3=In some lessons	
In lessons – Sciences	IS3G22D	4=In most lessons	
In lessons – Human sciences	IS3G22E	5=In every or almost every lesson	
In lessons – Creative arts	IS3G22F	9=Missing	
In lessons – Information technology	IS3G22G		
In lessons – Practical or vocational	IS3G22H		
In lessons – Other	IS3G22I		
During lessons – Word processing	IS3G23A	1=Never	PCA
During lessons – Presentation software	IS3G23B	2=In some lessons	
During lessons – Computer-based resources	IS3G23C	3=In most lessons	
During lessons – Multimedia tools	IS3G23D	4=In every or almost every lesson	
During lessons – Concept mapping	IS3G23E	9=Missing	
During lessons – Data capture tools	IS3G23F		
During lessons – Simulations and modeling	IS3G23G		
During lessons – Computer resources	IS3G23H		
During lessons – Interactive digital resources	IS3G23I		
During lessons – Graphic design	IS3G23J		
During lessons – Learning management system	IS3G23K		
During lessons – Video conferencing	IS3G23L		
During lessons – Programming	IS3G23M		
ICT well – Edit digital images	IS3G24A	1=Very well	PCA
ICT well – Write or edit assignments	IS3G24B	2=Moderately well	
ICT well – Search for project info online	IS3G24C	3=I have never done this, but I could work out how to do	
ICT well – Build or edit a webpage	IS3G24D	4=I do not think I could do this	
ICT well – Change device settings	IS3G24E	9=Missing	
ICT well – Create multimedia presentations	IS3G24F		
ICT well – Upload content to an online profile	IS3G24G		
ICT well – Insert an image into a document	IS3G24H		
ICT well – Install a program or app	IS3G24I		
ICT well – Judge trustworthiness of info	IS3G24J		
ICT well – Use text-based programming	IS3G24K		
ICT well – Use visual coding	IS3G24L		
ICT well – Find original sources of info	IS3G24M		



Table E.1: Background variables used for conditioning (cont'd)

Variable	Name	Values	Regressor
ICT and society - ICT improves living conditions	IS3G25AA	1=Strongly agree	PCA
ICT and society - ICT helps understand the world	IS3G25AB	2=Agree	
ICT and society - ICT makes people isolated	IS3G25AC	3=Disagree	
ICT and society - More ICT means fewer jobs	IS3G25AD	4=Strongly disagree	
ICT and society - Too much time on ICT	IS3G25AE	9=Missing	
ICT and society - ICT is valuable	IS3G25AF		
ICT and society - ICT brings social benefits	IS3G25AG		
ICT and society - ICT may harm health	IS3G25AH		
ICT and school - Learn better with ICT	IS3G25BA	1=Strongly agree	PCA
ICT and school - ICT makes learning fun	IS3G25BB	2=Agree	
ICT and school - Want to study ICT	IS3G25BC	3=Disagree	
ICT and school - Want ICT in my job	IS3G25BD	4=Strongly disagree	
ICT and school - Want a job in programming	IS3G25BE	9=Missing	
ICT and school - Learning ICT helps me	IS3G25BF		
ICT and school - Students need ICT in school	IS3G25BG		
ICT and school - ICT helps get jobs	IS3G25BH		
ICT and school - Students need programming	IS3G25BI		
ICT and school - Must keep up with ICT	IS3G25BJ		
Using ICT - Use solutions for problems	IS3G26A	1=To a large extent	PCA
Using ICT - Break problems into steps	IS3G26B	2=To a moderate extent	
Using ICT - Make concept diagrams	IS3G26C	3=To a small extent	
Using ICT - Plan tasks step-by-step	IS3G26D	4=Not at all	
Using ICT - Find data patterns	IS3G26E	9=Missing	
Using ICT - Use simulations to learn	IS3G26F		
Using ICT - Create flow diagrams	IS3G26G		
Using ICT - Test programs for bugs	IS3G26H		
Using ICT - Use data to solve problems	IS3G26I		
Using ICT - Study computing and IT	IS3G27	1=Yes - as a specific subject 2=Yes - within my other subjects 3=No 9=Missing	PCA

Appendix F:

National item adjudication

The table lists the items that were excluded from scaling in the corresponding national data when drawing plausible values but are included in the public-use database. Exclusion reasons include country DIF (differential item functioning), low ICR (inter-coder reliability), and translation error.



Table F.1: National item adjudication

Country	Reason	Item
Austria	Country DIF	G03Z
Azerbaijan	Country DIF	G01Z, G06Z, G08Z, I01Z, P03Z, P06F, R12C, S06Z
Belgium	Country DIF	G09G, I07G
Chinese Taipei	Country DIF	C07F, G02Z, P06F, R08Z
Cyprus	Country DIF	H01Z, H07J, I07G, S08G
Denmark	Country DIF	R06B, R12C
Finland	Country DIF	G09G, I07F, P06G, S03Z
France	Country DIF	G09G
Germany	Low ICR	I06Z, I07H
Greece	Country DIF	G07Z, H07J, R10Z, R12B, R12C
Hungary	Country DIF	G09G, H07J, I07E, P05Z
	Low ICR	P06E
Italy	Country DIF	G09G
Kazakhstan	Country DIF	C07F, I07F, I07G, P06F, R03Z, R12C, S03Z
Korea, Republic of	Country DIF	G07Z, H06Z, H07J, R12C, T3A08Z
Kosovo	Country DIF	C06Z, G03Z, G08Z, G09D, H02Z, I07C, I07D, P03Z, P05Z
Latvia	Country DIF	C07F, H07J, R12C, S03Z
Netherlands	Country DIF	G09E, R12C
North-Rhine W. (Germany)	Low ICR	H07C, R12C
Oman	Country DIF	G01Z, R12C
Portugal	Country DIF	H07J, I07G
Romania	Country DIF	H02Z, I07E, S04A
	Low ICR	H07G, H07I, H07J, C07B, C07C, C07G, I07F, I07G, I07H, I07I, P06C, P06E, P06F, P06G, P06H, R12A, R12D, R12I, R12B, R12C, R12G, R12H, R12J, P02Z
Serbia	Country DIF	C07A, G03Z, G09G, H07A, I07E, I07F, I07G, P06H, R12C
Slovakia	Country DIF	C07F
Slovenia	Country DIF	G03Z
Sweden	Translation error	TS05Z
	Low ICR	I07H, P06E
United States	Country DIF	I02Z
	Low ICR	C07G, P02Z
Uruguay	Low ICR	C07A, C07C, C07F, G08Z, P06C, P06E, P06G, P06H

Appendix G:

Recoding and scoring rules for variables included in scaling

This appendix details the recoding and scoring rules that were used to create the scoring variables included in scaling for the four released modules (two CIL and two CT modules). Scoring variables contain data about students' codified responses to tasks. In some cases a single scoring variable contains data corresponding directly to an evaluation of the quality of responses to a given task. The score categories in these variables may be used directly in scaling or may need to be recoded before the variable can be included in scaling. Other scoring variables are derived by combining information from more than one component variable, each of which captures specific details about how and/or how well a student executed a task. After the data from component variables have been combined, the resultant scoring variable may further need to be recoded before it can be included in scaling.

The scoring and recoding rules for the CIL modules, Breathing and School trip, are presented in [Table G.1](#) and [Table G.2](#), respectively. The rules for the CT modules, Automated bus and Farm drone, are shown in [Table G.3](#) and [Table G.4](#), respectively. Only scoring variables that resulted from combining variables or recoding are listed.

G.1 Breathing

Table G.1: Recoding rules for CIL module Breathing

Scoring variable	Original variable	Recoding rule
C1H02Z_S	C1H02ZA_P	$C1H02Z\_S = \begin{cases} 1 & \text{if } C1H02ZA\_P = 2 \\ 0 & \text{if } C1H02ZA\_P \in \{0, 1\} \end{cases}$

G.2 School trip

Table G.2: Scoring and recoding rules for CIL module School trip

Scoring variable	Original variable(s)	Scoring/Recoding rule
C1S02Z_S	C1S02ZA_P	$C1S02Z\_S = \begin{cases} 1 & \text{if } C1S02ZA\_P \in \{1, 2\} \\ 0 & \text{if } C1S02ZA\_P = 0 \end{cases}$
C1S06Z_S	C1S06ZMA_P, C1S06ZMB_P	$C1S06Z\_S = \begin{cases} 1 & \text{if } C1S06ZMA\_P = 1 \text{ and } C1S06ZMB\_P = 2 \\ 0 & \text{Otherwise} \end{cases}$
C1S08D_S	C1S08DL_P	$C1S08D\_S = \begin{cases} 2 & \text{if } C1S08DL\_P = 3 \\ 1 & \text{if } C1S08DL\_P \in \{1, 2\} \\ 0 & \text{if } C1S08DL\_P = 0 \end{cases}$



G.3 Automated bus

Table G.3: Scoring rubric for CT module Automated bus

Scoring variable	Original variable(s)	Scoring/Recoding rule
T2A01Z_S	T2A01ZA_P	$T2A01Z\_S = \begin{cases} 2 & \text{if } T2A01ZA\_P = 7 \\ 1 & \text{if } T2A01ZA\_P \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{if } T2A01ZA\_P = 0 \end{cases}$
T2A02Z_S	T2A02AA_P, T2A02BA_P, T2A02CA_P	$T2A02Z\_S = \begin{cases} 3 & \text{if } T2A02AA\_P = 11 \text{ and } T2A02BA\_P = 11 \text{ and } T2A02CA\_P = 2 \\ 2 & \text{if } T2A02AA\_P \in \{7, 8, 9, 10, 11\} \text{ and } \\ & T2A02BA\_P \leq (T2A02AA\_P + 3) \text{ and } T2A02CA\_P \in \{1, 2\} \\ 1 & T2A02AA\_P \in \{6, 7, 8, 9, 10, 11\} \text{ and } T2A02BA\_P \leq 17 \\ 0 & \text{Otherwise} \end{cases}$
T2A04Z_S	T2A04ZM_P	$T2A04Z\_S = \begin{cases} 1 & \text{if } T2A04ZM\_P = 1 \\ 0 & \text{if } T2A04ZM\_P \in \{2, 3, 4\} \end{cases}$
T3A05Z_S	T3A05ZA_P	$T3A05Z\_S = \begin{cases} 2 & \text{if } T3A05ZA\_P = 3 \\ 1 & \text{if } T3A05ZA\_P \in \{1, 2\} \\ 0 & \text{if } T3A05ZA\_P = 0 \end{cases}$
T3A06Z_S	T3A06AA_P, T3A06BA_P	$T3A06Z\_S = \begin{cases} 3 & \text{if } T3A06AA\_P = 4 \text{ and } T3A06BA\_P = 4 \\ 2 & \text{if } (T3A06AA\_P = 3 \text{ and } T3A06BA\_P = 5) \text{ or } \\ & (T3A06AA\_P = 2 \text{ and } T3A06BA\_P = 6) \text{ or } \\ & (T3A06AA\_P = 1 \text{ and } T3A06BA\_P = 7) \\ 1 & \text{if } (T3A06AA\_P = 4 \text{ and } T3A06BA\_P \in \{1, 2, 3\}) \text{ or } \\ & (T3A06AA\_P = 3 \text{ and } T3A06BA\_P \in \{1, 2, 3, 4\}) \text{ or } \\ & (T3A06AA\_P = 2 \text{ and } T3A06BA\_P \in \{1, 2, 3, 4, 5\}) \text{ or } \\ & (T3A06AA\_P = 1 \text{ and } T3A06BA\_P \in \{1, 2, 3, 4, 5, 6\}) \\ 0 & \text{Otherwise} \end{cases}$



G.4 Farm drone

Table G.4: Scoring rubric for CT module Farm drone

Scoring variable	Original variable(s)	Scoring/Recoding rule
T2F01E_S	T2F01AA_P, T2F01CA_P	$T2F01E\_S = \begin{cases} 3 & \text{if } T2F01AA\_P = 1 \text{ and } T2F01CA\_P = 1 \\ 2 & \text{if } T2F01AA\_P = 5 \text{ and } T2F01CA\_P = 1 \\ 1 & \text{if } T2F01CA\_P = 1 \\ 0 & \text{Otherwise} \end{cases}$
T2F02L_S	T2F02AA_P, T2F02ZY_P	$T2F02L\_S = \begin{cases} 2 & \text{if } T2F02ZY\_P = (X5, Y7) \text{ and } T2F02AA\_P = 4 \\ 1 & \text{if } T2F02ZY\_P \in \{(X5, Y7), (X6, Y7), (X6, Y8), (X5, Y8), (X4, Y8), \\ & (X4, Y7), (X4, Y6), (X5, Y6), (X6, Y6)\} \text{ and } T2F02AA\_P \leq 10 \\ 0 & \text{Otherwise} \end{cases}$
T2F03TEC_S	T2F03AA_P, T2F03WA_P, T2F03XA_P	$T2F03\_E = \begin{cases} 3 & \text{if } T2F03AA\_P = 3 \\ 2 & \text{if } T2F03AA\_P \in \{4, 5, 6, 7, 8\} \\ 1 & \text{if } T2F03AA\_P \in \{9, 10, 11, 12\} \\ 0 & \text{Otherwise} \end{cases}$ $T2F03\_T = \begin{cases} 4 & \text{if } T2F03WA\_P = 1 \text{ and } T2F03XA\_P = 8 \\ 3 & \text{if } T2F03WA\_P = 0 \text{ and } T2F03XA\_P = 8 \\ 2 & \text{if } T2F03WA\_P = 1 \text{ and } T2F03XA\_P \geq 4 \\ 1 & \text{if } T2F03WA\_P = 0 \text{ and } T2F03XA\_P \geq 4 \\ 0 & \text{Otherwise} \end{cases}$ $T2F03\_T2 = \begin{cases} 2 & \text{if } T2F03\_T = 4 \\ 1 & \text{if } T2F03\_T \in \{2, 3\} \\ 0 & \text{if } T2F03\_T \in \{0, 1\} \end{cases}$ $T2F03TEC\_S = \begin{cases} 3 & \text{if } T2F03\_T2 = 2 \text{ and } T2F03\_E = 3 \\ 1 & \text{if } T2F03\_T2 = 2 \text{ and } T2F03\_E = 0 \\ T2F03\_T2 & \text{Otherwise} \end{cases}$
T3F04TEC_S	T3F04AA_P, T3F04WA_P, T3F04XA_P	$T3F04\_E = \begin{cases} 3 & \text{if } T3F04AA\_P = 6 \\ 2 & \text{if } T3F04AA\_P \in \{7, 8, 9, 10\} \\ 1 & \text{if } T3F04AA\_P \in \{11, 12, 13, 14, 15\} \\ 0 & \text{Otherwise} \end{cases}$ $T3F04\_T = \begin{cases} 4 & \text{if } T3F04WA\_P = 1 \text{ and } T3F04XA\_P = 8 \\ 3 & \text{if } T3F04WA\_P = 0 \text{ and } T3F04XA\_P = 8 \\ 2 & \text{if } T3F04WA\_P = 1 \text{ and } T3F04XA\_P \geq 4 \\ 1 & \text{if } T3F04WA\_P = 0 \text{ and } T3F04XA\_P \geq 4 \\ 0 & \text{Otherwise} \end{cases}$ $T3F04\_T2 = \begin{cases} 2 & \text{if } T3F04\_T = 4 \\ 1 & \text{if } T3F04\_T \in \{2, 3\} \\ 0 & \text{if } T3F04\_T \in \{0, 1\} \end{cases}$ $T3F04TEC\_S = \begin{cases} 3 & \text{if } T3F04\_T2 = 2 \text{ and } T3F04\_E = 3 \\ 1 & \text{if } T3F04\_T2 = 2 \text{ and } T3F04\_E = 0 \\ T3F04\_T2 & \text{Otherwise} \end{cases}$
T2F05TEC_S	T2F05AA_P, T2F05WA_P, T2F05XA_P	$T2F05\_E = \begin{cases} 2 & \text{if } T2F05AA\_P \in \{12, 13, 14\} \\ 1 & \text{if } T2F05AA\_P \in \{15, 16, 17, 18, 19, 20, 21, 22\} \\ 0 & \text{Otherwise} \end{cases}$ $T2F05\_T = \begin{cases} 6 & \text{if } T2F05WA\_P = 1 \text{ and } T2F05XA\_P = 16 \\ 5 & \text{if } T2F05WA\_P = 0 \text{ and } T2F05XA\_P = 16 \\ 4 & \text{if } T2F05WA\_P = 1 \text{ and } T2F05XA\_P \geq 12 \\ 3 & \text{if } T2F05WA\_P = 0 \text{ and } T2F05XA\_P \geq 12 \\ 2 & \text{if } T2F05WA\_P = 1 \text{ and } T2F05XA\_P \geq 8 \\ 1 & \text{if } T2F05WA\_P = 0 \text{ and } T2F05XA\_P \geq 8 \\ 0 & \text{Otherwise} \end{cases}$ $T2F05\_T2 = \begin{cases} 2 & \text{if } T2F05\_T = 6 \\ 1 & \text{if } T2F05\_T \in \{2, 3, 4, 5\} \\ 0 & \text{if } T2F05\_T \in \{0, 1\} \end{cases}$ $T2F05TEC\_S = \begin{cases} 3 & \text{if } T2F05\_T2 = 2 \text{ and } T2F05\_E = 2 \\ 1 & \text{if } T2F05\_T2 = 2 \text{ and } T2F05\_E = 0 \\ T2F05\_T2 & \text{Otherwise} \end{cases}$



Table G.4: Scoring rubric for CT module Farm drone (Continued)

Variable	Component(s)	Scoring Rule
T3F06TEC_S	T3F06AA_P, T3F06WA_P, T3F06XA_P, T3F06MA_P, T3F06NA_P, T3F06OA_P, T3F06PA_P, T3F06QA_P, T3F06RA_P	$T3F06\_E = \begin{cases} 3 & \text{if } T3F06AA\_P = 5 \\ 2 & \text{if } T3F06AA\_P \in \{6, 7\} \\ 1 & \text{if } T3F06AA\_P \in \{8, 9, 10, 11, 12\} \\ 0 & \text{Otherwise} \end{cases}$ $T3F06\_T = \begin{cases} 6 & \text{if } T3F06WA\_P = 1 \text{ and } T3F06XA\_P = 12 \\ 5 & \text{if } T3F06WA\_P = 0 \text{ and } T3F06XA\_P = 12 \\ 4 & \text{if } T3F06MA\_P = 2 \text{ and } T3F06NA\_P = 0 \text{ and } T3F06OA\_P = 0 \text{ and} \\ & T3F06PA\_P = 2 \text{ and } T3F06QA\_P = 0 \text{ and } T3F06RA\_P = 2 \text{ and} \\ & T3F06XA\_P = 6 \text{ and } T3F06WA\_P = 1 \\ 3 & \text{if } T3F06WA\_P = 1 \text{ and } T3F06XA\_P \geq 8 \\ 2 & \text{if } T3F06WA\_P = 0 \text{ and } T3F06XA\_P \geq 8 \\ 1 & \text{if } T3F06XA\_P \geq 6 \\ 0 & \text{Otherwise} \end{cases}$ $T3F06\_T2 = \begin{cases} 2 & \text{if } T3F06\_T = 6 \\ 1 & \text{if } T3F06\_T \in \{3, 4, 5\} \\ 0 & \text{if } T3F06\_T \in \{0, 1, 2\} \end{cases}$ $T3F06TEC\_S = \begin{cases} 3 & \text{if } T3F06\_T2 = 2 \text{ and } T3F06\_E = 3 \\ 1 & \text{if } T3F06\_T2 = 2 \text{ and } T3F06\_E = 0 \\ T3F06\_T2 & \text{Otherwise} \end{cases}$
T2F07TEC_S	T2F07AA_P, T2F07WA_P, T2F07XA_P, T2F07MA_P, T2F07NA_P, T2F07OA_P, T2F07PA_P, T2F07QA_P, T2F07RA_P	$T2F07\_E = \begin{cases} 3 & \text{if } T2F07AA = 12 \\ 2 & \text{if } T2F07AA \in \{13, 14, 15, 16, 17, 18, 19\} \\ 1 & \text{if } T2F07AA \in \{20, 21, 22, 23, 24\} \\ 0 & \text{Otherwise} \end{cases}$ $T2F07\_T = \begin{cases} 6 & \text{if } T2F07WA\_P = 1 \text{ and } T2F07XA\_P = 24 \\ 5 & \text{if } T2F07WA\_P = 0 \text{ and } T2F07XA\_P = 24 \\ 4 & \text{if } T2F07MA\_P = 4 \text{ and } T2F07NA\_P = 0 \text{ and } T2F07OA\_P = 0 \text{ and} \\ & T2F07PA\_P = 4 \text{ and } T2F07QA\_P = 0 \text{ and } T2F07RA\_P = 4 \text{ and} \\ & T2F07XA\_P = 12 \text{ and } T2F07WA\_P = 1 \\ 3 & \text{if } T2F07WA\_P = 1 \text{ and } T2F07XA\_P \geq 16 \\ 2 & \text{if } T2F07WA\_P = 0 \text{ and } T2F07XA\_P \geq 16 \\ 1 & \text{if } T2F07XA\_P \geq 12 \\ 0 & \text{Otherwise} \end{cases}$ $T2F07\_T2 = \begin{cases} 3 & \text{if } T2F07\_T \in \{5, 6\} \\ 2 & \text{if } T2F07\_T \in \{3, 4\} \\ 1 & \text{if } T2F07\_T = 2 \\ 0 & \text{if } T2F07\_T \in \{0, 1\} \end{cases}$ $T2F07 = \begin{cases} 4 & \text{if } T2F07\_T2 = 3 \text{ and } T2F07\_E \in \{2, 3\} \\ 2 & \text{if } T2F07\_T2 = 3 \text{ and } T2F07\_E = 0 \\ T2F07\_T2 & \text{Otherwise} \end{cases}$ $T2F07TEC\_S = \begin{cases} 3 & \text{if } T2F07 = 4 \\ 2 & \text{if } T2F07 = 3 \\ 1 & \text{if } T2F07 \in \{1, 2\} \\ 0 & \text{if } T2F07 = 0 \end{cases}$



Table G.4: Scoring rubric for CT module Farm drone (Continued)

Variable	Component(s)	Scoring Rule
T2F08TEC_S	T2F08AA_P, T2F08WA_P, T2F08XA_P, T2F08DA_P, T2F08PA_P, T2F08QA_P, T2F08RA_P	$T2F08\_E = \begin{cases} 4 & \text{if } T2F08AA\_P \in \{12, 13, 14\} \\ 3 & \text{if } T2F08AA\_P = 15 \\ 2 & \text{if } T2F08AA\_P \in \{16, 17, 18, 19, 20, 21, 22\} \\ 1 & \text{if } T2F08AA\_P \in \{23, 24, \dots, 42\} \\ 0 & \text{Otherwise} \end{cases}$ $T2F08\_T = \begin{cases} 7 & \text{if } T2F08WA\_P = 1 \text{ and } T2F08XA\_P = 32 \\ 6 & \text{if } T2F08WA\_P = 0 \text{ and } T2F08XA\_P = 32 \\ 5 & \text{if } T2F08WA\_P = 1 \text{ and } T2F08XA\_P \geq 24 \\ 4 & \text{if } T2F08WA\_P = 0 \text{ and } T2F08XA\_P \geq 24 \\ 3 & \text{if } T2F08WA\_P = 1 \text{ and } T2F08XA\_P \geq 14 \\ 2 & \text{if } T2F08WA\_P = 0 \text{ and } T2F08XA\_P \geq 14 \\ 1 & \text{if } T2F08AA\_P = 15 \text{ and } T2F08DA\_P = 2 \text{ and } \\ & \quad T2F08PA\_P = 14 \text{ and } T2F08QA\_P = 0 \text{ and } \\ & \quad T2F08RA\_P = 0 \text{ and } T2F08XA\_P = 28 \text{ and } \\ & \quad T2F08WA\_P = 0 \\ 0 & \text{Otherwise} \end{cases}$ $T2F08\_T2 = \begin{cases} 2 & \text{if } T2F08\_T = 7 \\ 1 & \text{if } T2F08\_T \in \{3, 5\} \\ 0 & \text{if } T2F08\_T \in \{1, 2, 4, 6\} \end{cases}$ $T2F08 = \begin{cases} 3 & \text{if } T2F08\_T2 = 2 \text{ and } T2F08\_E \in \{3, 4\} \\ 1 & \text{if } T2F08\_T2 = 2 \text{ and } T2F08\_E = 0 \\ T2F08\_T2 & \text{Otherwise} \end{cases}$ $T2F08TEC\_S = \begin{cases} 2 & \text{if } T2F08 \in \{2, 3\} \\ 1 & \text{if } T2F08 = 1 \\ 0 & \text{if } T2F08 = 0 \end{cases}$
T2F09T_S	T2F09PA_P	$T2F09T = \begin{cases} 5 & \text{if } T2F09PA\_P = 16 \\ 4 & \text{if } T2F09PA\_P \in \{12, 13, 14, 15\} \\ 3 & \text{if } T2F09PA\_P \in \{9, 10, 11\} \\ 2 & \text{if } T2F09PA\_P = 8 \\ 1 & \text{if } T2F09PA\_P \in \{5, 6, 7\} \\ 0 & \text{Otherwise} \end{cases}$ $T2F09T\_S = \begin{cases} 2 & \text{if } T2F09T = 5 \\ 1 & \text{if } T2F09T = 4 \\ 0 & \text{if } T2F09T \in \{0, 1, 2, 3\} \end{cases}$

Appendix H:

Details on questionnaire scale results

Table H.1: Reliabilities by country for scales from student questionnaire (part 1)

Country	S_LRNINTS	S_LRNINTO	S_ACMULT	S_GENEFF	S_SPECEFF
Austria	0.783	0.766	0.812	0.865	0.833
Azerbaijan	0.863	0.755	0.825	0.905	0.791
Belgium (Flemish)	0.780	0.793	0.880	0.881	0.757
Bosnia and Herzegovina	0.873	0.800	0.868	0.932	0.832
†Chile	0.856	0.820	0.868	0.928	0.846
Chinese Taipei	0.881	0.900	0.922	0.940	0.829
Croatia	0.877	0.833	0.893	0.960	0.850
Cyprus	0.875	0.836	0.884	0.934	0.834
Czech Republic	0.861	0.809	0.812	0.900	0.783
Denmark	0.728	0.751	0.890	0.877	0.831
Finland	0.828	0.831	0.911	0.920	0.862
France	0.784	0.774	0.844	0.889	0.744
Germany	0.783	0.754	0.848	0.874	0.821
Greece	0.872	0.791	0.853	0.904	0.813
Hungary	0.853	0.815	0.836	0.889	0.817
Italy	0.862	0.755	0.847	0.864	0.773
Kazakhstan	0.812	0.779	0.854	0.913	0.779
Korea, Republic of	0.884	0.919	0.873	0.943	0.812
Kosovo	0.795	0.714	0.800	0.879	0.729
Latvia	0.847	0.832	0.863	0.912	0.780
Luxembourg	0.781	0.779	0.832	0.900	0.807
Malta	0.882	0.826	0.898	0.937	0.825
†Netherlands	0.802	0.832	0.860	0.915	0.827
Norway	0.778	0.775	0.910	0.913	0.802
Oman	0.762	0.702	0.766	0.888	0.739
Portugal	0.863	0.809	0.854	0.923	0.806
†Romania	0.886	0.855	0.898	0.934	0.843
Serbia	0.896	0.841	0.884	0.940	0.822
Slovak Republic	0.878	0.856	0.815	0.904	0.807
Slovenia	0.850	0.830	0.865	0.927	0.849
Spain	0.822	0.787	0.841	0.887	0.783
Sweden	0.799	0.814	0.903	0.914	0.799
†United States	0.756	0.785	0.891	0.921	0.798
Uruguay	0.726	0.724	0.854	0.888	0.737
†North Rhine-Westphalia, Germany	0.766	0.752	0.856	0.876	0.841
International Average	0.844	0.809	0.866	0.910	0.810

† These countries are excluded from the international calibration.



Table H.2: Reliabilities by country for scales from student questionnaire (part 2)

Country	S_ICTPOSG	S_ICTNEG	S_ICTFUT	S_CTCLS	S_LRNSAFE
Austria	0.795	0.737	0.871	0.912	0.862
Azerbaijan	0.829	0.731	0.866	0.917	0.750
Belgium (Flemish)	0.752	0.673	0.888	0.892	0.793
Bosnia and Herzegovina	0.831	0.776	0.897	0.935	0.793
†Chile	0.863	0.809	0.891	0.930	0.869
Chinese Taipei	0.936	0.810	0.894	0.944	0.918
Croatia	0.889	0.839	0.913	0.943	0.888
Cyprus	0.859	0.805	0.888	0.925	0.886
Czech Republic	0.784	0.739	0.871	0.917	0.851
Denmark	0.799	0.731	0.857	0.870	0.835
Finland	0.878	0.781	0.906	0.925	0.893
France	0.804	0.718	0.853	0.890	0.814
Germany	0.771	0.725	0.869	0.904	0.834
Greece	0.803	0.753	0.841	0.903	0.844
Hungary	0.826	0.727	0.882	0.907	0.860
Italy	0.797	0.708	0.853	0.891	0.801
Kazakhstan	0.827	0.685	0.860	0.911	0.819
Korea, Republic of	0.925	0.774	0.871	0.940	0.885
Kosovo	0.767	0.721	0.855	0.878	0.698
Latvia	0.830	0.702	0.872	0.905	0.844
Luxembourg	0.777	0.720	0.871	0.900	0.819
Malta	0.866	0.761	0.921	0.926	0.854
†Netherlands	0.853	0.756	0.893	0.912	0.842
Norway	0.857	0.778	0.881	0.924	0.880
Oman	0.754	0.669	0.790	0.863	0.660
Portugal	0.845	0.776	0.887	0.937	0.842
†Romania	0.866	0.825	0.900	0.936	0.859
Serbia	0.846	0.802	0.900	0.927	0.873
Slovak Republic	0.817	0.699	0.862	0.927	0.850
Slovenia	0.851	0.832	0.904	0.942	0.885
Spain	0.842	0.746	0.863	0.899	0.835
Sweden	0.889	0.805	0.886	0.924	0.895
†United States	0.827	0.773	0.865	0.890	0.832
Uruguay	0.765	0.745	0.848	0.882	0.785
†North Rhine-Westphalia, Germany	0.780	0.731	0.884	0.913	0.848
International Average	0.830	0.751	0.879	0.919	0.845

† These countries are excluded from the international calibration.



Table H.3: Reliabilities by country for scales from student questionnaire (part 3)

Country	S_GENCLASS	S_SPECLASS	S_ICTPOSS
Austria	0.738	0.906	0.751
Azerbaijan	0.875	0.922	0.830
Belgium (Flemish)	0.668	0.882	0.745
Bosnia and Herzegovina	0.846	0.943	0.826
†Chile	0.826	0.945	0.852
Chinese Taipei	0.868	0.908	0.851
Croatia	0.855	0.928	0.841
Cyprus	0.854	0.939	0.847
Czech Republic	0.784	0.866	0.783
Denmark	0.663	0.903	0.705
Finland	0.708	0.917	0.786
France	0.758	0.898	0.784
Germany	0.692	0.897	0.763
Greece	0.829	0.921	0.791
Hungary	0.796	0.906	0.737
Italy	0.765	0.872	0.753
Kazakhstan	0.862	0.905	0.839
Korea, Republic of	0.852	0.936	0.839
Kosovo	0.763	0.904	0.824
Latvia	0.777	0.871	0.755
Luxembourg	0.735	0.914	0.767
Malta	0.795	0.924	0.857
†Netherlands	0.738	0.925	0.802
Norway	0.752	0.836	0.824
Oman	0.727	0.869	0.773
Portugal	0.787	0.920	0.843
†Romania	0.841	0.938	0.878
Serbia	0.811	0.913	0.831
Slovak Republic	0.831	0.896	0.779
Slovenia	0.840	0.936	0.846
Spain	0.757	0.906	0.815
Sweden	0.686	0.926	0.804
†United States	0.722	0.916	0.770
Uruguay	0.695	0.883	0.790
†North Rhine-Westphalia, Germany	0.725	0.910	0.755
International Average	0.803	0.912	0.800

† These countries are excluded from the international calibration.



Table H.4: Reliabilities by country for scales from teacher questionnaire (part 1)

Country	T_EMPITE	T_COLICT	T_RESRC	T_VWNEG	T_VWPOS
†Austria	0.964	0.920	0.898	0.888	0.865
Azerbaijan	0.981	0.949	0.939	0.893	0.909
†Belgium (Flemish)	0.962	0.913	0.901	0.863	0.839
†Bosnia and Herzegovina	0.989	0.967	0.915	0.916	0.889
†Chile	0.972	0.946	0.877	0.866	0.888
Chinese Taipei	0.982	0.962	0.935	0.914	0.922
Croatia	0.990	0.953	0.904	0.901	0.894
†Cyprus	0.98	0.965	0.931	0.907	0.902
Czech Republic	0.978	0.938	0.902	0.879	0.863
†Denmark	0.968	0.900	0.877	0.862	0.866
Finland	0.961	0.928	0.861	0.878	0.851
†France	0.961	0.941	0.878	0.885	0.873
†Germany	0.940	0.887	0.888	0.875	0.846
Greece	0.982	0.968	0.934	0.913	0.906
Hungary	0.987	0.954	0.902	0.900	0.901
Italy	0.968	0.946	0.881	0.914	0.905
Kazakhstan	0.980	0.924	0.929	0.864	0.902
Korea, Republic of	0.968	0.974	0.912	0.917	0.898
†Kosovo	0.941	0.933	0.899	0.854	0.913
Latvia	0.973	0.930	0.860	0.858	0.861
†Luxembourg	0.960	0.908	0.884	0.906	0.882
†Malta	0.960	0.953	0.845	0.884	0.861
†Netherlands	0.945	0.898	0.865	0.859	0.788
Norway	0.968	0.925	0.868	0.845	0.832
†Oman	0.942	0.948	0.912	0.884	0.912
Portugal	0.986	0.937	0.896	0.890	0.913
†Romania	0.945	0.968	0.941	0.904	0.930
Serbia	0.986	0.944	0.915	0.884	0.920
†Slovak Republic	0.979	0.942	0.900	0.877	0.872
Slovenia	0.980	0.938	0.864	0.888	0.853
Spain	0.977	0.947	0.897	0.903	0.905
†Sweden	0.969	0.928	0.859	0.867	0.878
†United States	0.980	0.960	0.922	0.918	0.909
†Uruguay	0.970	0.961	0.903	0.876	0.922
†North Rhine-Westphalia, Germany	0.946	0.905	0.879	0.877	0.859
International Average	0.982	0.947	0.903	0.888	0.890

† These countries are excluded from the international calibrations. – The items belonging to the scale are not administered in these countries.



Table H.5: Reliabilities by country for scales from teacher questionnaire (part 2)

Country	T_TEAPTC	T_TEAPIN	T_USETOOL	T_CODEMP	T_EPIEMC
†Austria	0.803	0.907	0.866	0.902	—
Azerbaijan	0.916	0.903	0.945	0.919	0.864
†Belgium (Flemish)	0.756	0.892	0.809	0.903	0.753
†Bosnia and Herzegovina	0.839	0.935	0.947	0.944	0.836
†Chile	0.830	0.932	0.917	0.937	0.809
Chinese Taipei	0.909	0.922	0.913	0.939	0.933
Croatia	0.860	0.917	0.895	0.921	0.773
†Cyprus	0.867	0.942	0.891	0.934	0.850
Czech Republic	0.814	0.917	0.877	0.906	—
†Denmark	0.746	0.898	—	0.915	0.650
Finland	0.795	0.912	0.776	0.904	—
†France	0.780	0.894	0.848	0.915	0.778
†Germany	0.770	0.899	0.849	—	—
Greece	0.864	0.937	0.907	0.924	0.808
Hungary	0.800	0.931	0.889	0.928	0.807
Italy	0.795	0.910	0.841	0.884	0.798
Kazakhstan	0.825	0.898	0.929	0.915	0.842
Korea, Republic of	0.832	0.921	0.921	0.928	0.916
†Kosovo	0.895	0.919	0.930	0.933	0.849
Latvia	0.765	0.892	0.870	0.921	0.725
†Luxembourg	0.755	0.901	0.830	0.924	—
†Malta	0.777	0.903	0.858	0.921	0.836
†Netherlands	0.782	0.889	0.860	0.916	0.702
Norway	0.746	0.820	0.825	0.897	0.721
†Oman	0.883	0.928	0.935	0.945	0.859
Portugal	0.854	0.920	0.864	0.906	0.804
†Romania	0.915	0.942	0.939	0.947	0.846
Serbia	0.847	0.934	0.914	0.919	0.856
†Slovak Republic	0.852	0.94	0.86	0.922	0.78
Slovenia	0.868	0.922	0.860	0.929	0.832
Spain	0.799	0.920	0.874	0.900	0.824
†Sweden	0.748	0.886	0.831	0.927	0.774
†United States	0.885	0.940	0.858	0.935	0.854
†Uruguay	0.834	0.912	0.890	0.909	0.814
†North Rhine-Westphalia, Germany	0.766	0.893	0.881	—	—
International Average	0.817	0.923	0.905	0.923	0.835

† These countries are excluded from the international calibrations. — The items belonging to the scale are not administered in these countries.



Table H.6: Reliabilities by country for scales from teacher questionnaire (part 3)

Country	T_ICTCLASB	T_ICTCLASA	T_EPICOG	T_ICTEFF	T_PROFLRN
†Austria	0.855	0.930	—	0.904	0.902
Azerbaijan	0.914	0.943	0.902	0.943	0.961
†Belgium (Flemish)	0.821	0.925	0.781	0.915	0.904
†Bosnia and Herzegovina	0.884	0.960	0.845	0.939	0.951
†Chile	0.874	0.931	0.830	0.944	0.952
Chinese Taipei	0.911	0.951	0.943	0.942	0.923
Croatia	0.863	0.928	0.835	0.946	0.935
†Cyprus	0.884	0.926	0.867	0.95	0.936
Czech Republic	0.861	0.919	—	0.926	0.923
†Denmark	0.789	0.911	0.660	0.895	0.929
Finland	0.807	0.892	—	0.922	0.923
†France	0.773	0.889	0.745	0.914	0.927
†Germany	0.808	0.897	—	0.899	0.883
Greece	0.902	0.946	0.835	0.946	0.932
Hungary	0.893	0.926	0.827	0.924	0.941
Italy	0.835	0.905	0.779	0.925	0.932
Kazakhstan	0.879	0.934	0.900	0.940	0.914
Korea, Republic of	0.903	0.945	0.927	0.949	0.939
†Kosovo	0.890	0.937	0.897	0.928	0.946
Latvia	0.827	0.911	0.824	0.926	0.897
†Luxembourg	0.839	0.890	—	0.903	0.891
†Malta	0.860	0.911	0.844	0.909	0.917
†Netherlands	0.844	0.926	0.735	0.913	0.904
Norway	0.794	0.896	0.729	0.917	0.928
†Oman	0.868	0.948	0.908	0.930	0.925
Portugal	0.878	0.925	0.796	0.941	0.929
†Romania	0.857	0.945	0.866	0.934	0.957
Serbia	0.890	0.947	0.879	0.942	0.945
†Slovak Republic	0.880	0.925	0.815	0.922	0.922
Slovenia	0.853	0.940	0.850	0.918	0.900
Spain	0.859	0.913	0.810	0.937	0.918
†Sweden	0.791	0.893	0.773	0.910	0.910
†United States	0.848	0.921	0.867	0.961	0.937
†Uruguay	0.847	0.918	0.812	0.925	0.926
†North Rhine-Westphalia, Germany	0.784	0.909	—	0.908	0.888
International Average	0.884	0.942	0.862	0.931	0.934

† These countries are excluded from the international calibrations. — The items belonging to the scale are not administered in these countries.



Table H.7: Reliabilities by country for scales from teacher questionnaire (part 4)

Country	T_PROFNEED	T_USEUTIL	T_ICTEMP	T_EPICON	T_TEAPEX
†Austria	0.920	0.805	0.955	—	0.804
Azerbaijan	0.978	0.931	0.969	0.887	0.894
†Belgium (Flemish)	0.917	0.781	0.954	0.777	0.863
†Bosnia and Herzegovina	0.976	0.895	0.975	0.882	0.861
†Chile	0.973	0.864	0.956	0.833	0.843
Chinese Taipei	0.974	0.865	0.966	0.903	0.833
Croatia	0.965	0.839	0.966	0.835	0.856
†Cyprus	0.962	0.825	0.970	0.877	0.882
Czech Republic	0.940	0.795	0.948	—	0.748
†Denmark	0.932	—	0.940	0.635	0.818
Finland	0.928	0.737	0.962	—	0.727
†France	0.932	0.749	0.957	0.744	0.695
†Germany	0.916	0.805	0.942	—	0.804
Greece	0.953	0.879	0.971	0.856	0.880
Hungary	0.956	0.881	0.962	0.789	0.837
Italy	0.951	0.819	0.946	0.764	0.833
Kazakhstan	0.948	0.881	0.950	0.880	0.833
Korea, Republic of	0.966	0.861	0.962	0.884	0.890
†Kosovo	0.970	0.898	0.955	0.873	0.865
Latvia	0.924	0.830	0.955	0.760	0.840
†Luxembourg	0.916	0.777	0.958	—	0.744
†Malta	0.931	0.761	0.954	0.864	0.842
†Netherlands	0.905	0.731	0.956	0.687	0.873
Norway	0.933	0.755	0.942	0.714	0.810
†Oman	0.958	0.872	0.953	0.882	0.891
Portugal	0.947	0.801	0.954	0.773	0.850
†Romania	0.973	0.899	0.966	0.843	0.895
Serbia	0.97	0.85	0.959	0.874	0.892
†Slovak Republic	0.946	0.851	0.955	0.815	0.801
Slovenia	0.946	0.820	0.961	0.848	0.855
Spain	0.939	0.818	0.949	0.787	0.819
†Sweden	0.925	0.765	0.958	0.742	0.799
†United States	0.952	0.811	0.973	0.838	0.861
†Uruguay	0.957	0.849	0.945	0.808	0.790
†North Rhine-Westphalia, Germany	0.916	0.786	0.942	—	0.788
International Average	0.956	0.841	0.959	0.841	0.857

† These countries are excluded from the international calibrations. — The items belonging to the scale are not administered in these countries.



Table H.8: Reliabilities for the pooled international dataset for principal and ICT coordinator scales

Scale	Cronbach's alpha
C_HINRES	0.906
C_ICTRES	0.777
C_RESTOOL	0.860
C_RESUTIL	0.852
C_HINPED	0.884
P_VWICT	0.940
P_EXPLRN	0.914
P_EXPTCH	0.861
P_PRIORS	0.926
P_ICTUSE	0.819
P_ICTCOM	0.844
P_CIMNEG	0.750
P_CIMPOS	0.909
P_CWORK	0.852
P_CONEG	0.836



Table H.9: Proportion of explained variance by the first principal component by country for scales from student questionnaire (part 1)

Country	S_LRNINTS	S_LRNINTO	S_ACMULT	S_GENEFF	S_SPECEFF
Austria	48.0	46.1	52.8	45.5	75.1
Azerbaijan	59.4	45.0	53.4	54.0	70.7
Belgium (Flemish)	47.7	49.2	63.2	48.6	67.4
Bosnia and Herzegovina	61.2	50.1	60.4	62.2	75.0
† Chile	58.1	52.8	60.2	61.1	76.7
Chinese Taipei	62.7	66.8	72.4	65.3	74.8
Croatia	62.0	54.5	65.2	73.9	77.1
Cyprus	61.7	55.0	63.5	63.2	75.2
Czech Republic	59.0	51.1	52.4	53.0	69.8
Denmark	43.2	44.8	64.7	48.3	74.9
Finland	54.0	54.4	69.6	58.3	78.5
France	48.2	47.1	56.7	50.3	66.4
Germany	48.1	44.9	57.5	47.0	73.8
Greece	61.1	48.9	57.9	54.2	72.9
Hungary	57.7	51.9	55.2	50.4	73.3
Italy	59.2	45.0	57.3	45.3	69.1
Kazakhstan	51.5	47.5	57.8	56.3	69.4
Korea, Republic of	63.4	71.2	61.3	66.6	72.8
Kosovo	49.8	41.8	50.1	48.3	65.1
Latvia	56.7	54.3	59.5	56.3	69.5
Luxembourg	47.8	47.6	54.7	53.0	72.3
Malta	62.9	53.5	66.6	64.8	74.3
† Netherlands	50.4	54.5	59.3	57.0	74.5
Norway	47.5	47.1	69.2	57.1	71.9
Oman	45.7	40.2	46.3	50.4	66.0
Portugal	59.3	51.2	58.1	59.4	72.3
† Romania	63.8	57.9	66.4	63.2	76.3
Serbia	65.8	55.7	63.5	65.1	73.8
Slovak Republic	62.2	58.1	52.3	54.0	72.2
Slovenia	57.3	54.2	60.1	60.5	77.0
Spain	52.9	48.4	56.1	50.0	69.9
Sweden	50.0	51.9	67.8	57.0	71.6
† United States	45.2	48.2	65.0	59.6	71.5
Uruguay	42.3	42.1	58.3	50.3	65.8
† North Rhine-Westphalia, Germany	46.2	44.8	58.8	47.5	76.0
International Average	56.3	51.2	60.0	55.6	72.6

Note: † These countries are excluded from the pooled international dataset.



Table H.10: Proportion of explained variance by the first principal component by country for scales from student questionnaire (part 2)

Country	S_ICTPOSG	S_ICTNEG	S_ICTFUT	S_CTCLS	S_LRNSAFE
Austria	62.0	56.0	66.3	58.9	70.8
Azerbaijan	66.1	55.4	65.1	60.2	57.5
Belgium (Flemish)	57.4	50.6	69.7	54.0	61.9
Bosnia and Herzegovina	66.5	59.8	70.9	66.0	62.3
†Chile	70.8	63.7	69.8	64.3	71.9
Chinese Taipei	83.9	63.8	70.5	69.2	80.3
Croatia	75.0	67.5	74.4	68.8	74.9
Cyprus	70.3	63.1	69.2	62.4	74.6
Czech Republic	60.7	56.3	66.6	60.2	69.3
Denmark	62.4	55.5	64.2	49.4	67.0
Finland	73.3	60.5	72.9	62.9	75.7
France	63.0	54.3	63.2	53.2	64.5
Germany	59.2	55.2	66.4	56.7	66.9
Greece	62.8	57.5	61.4	56.2	68.2
Hungary	65.8	55.2	68.2	57.5	70.4
Italy	62.1	53.6	63.1	53.5	62.8
Kazakhstan	65.8	51.4	64.3	58.5	64.9
Korea, Republic of	81.7	59.7	66.4	67.7	74.6
Kosovo	58.9	54.4	63.5	50.6	52.8
Latvia	66.2	52.8	66.5	56.9	68.3
Luxembourg	60.0	54.6	66.3	55.8	65.1
Malta	71.4	58.3	76.2	63.0	69.7
†Netherlands	69.4	57.8	70.3	59.1	68.0
Norway	70.0	60.1	68.0	62.3	73.7
Oman	57.7	50.3	54.5	47.8	49.7
Portugal	68.3	59.9	69.1	66.7	68.0
†Romania	71.3	65.5	71.6	66.3	70.3
Serbia	68.4	62.8	71.5	63.1	72.5
Slovak Republic	64.6	52.8	64.8	63.3	69.1
Slovenia	69.2	66.7	72.4	68.5	74.5
Spain	67.9	56.9	64.8	55.6	67.1
Sweden	75.0	63.3	69.2	62.5	76.1
†United States	65.9	59.6	65.1	53.3	66.7
Uruguay	58.7	56.8	62.6	51.5	61.2
†North Rhine-Westphalia, Germany	60.3	55.7	68.8	59.1	68.8
International Average	66.2	57.3	67.6	60.6	68.4

Note: † These countries are excluded from the pooled international dataset.



Table H.11: Proportion of explained variance by the first principal component by country for scales from student questionnaire (part 3)

Country	S_ICTPOSG	S_ICTNEG	S_ICTFUT	S_CTCLS	S_LRNSAFE
Austria	62.0	56.0	66.3	58.9	70.8
Azerbaijan	66.1	55.4	65.1	60.2	57.5
Belgium (Flemish)	57.4	50.6	69.7	54.0	61.9
Bosnia and Herzegovina	66.5	59.8	70.9	66.0	62.3
† Chile	70.8	63.7	69.8	64.3	71.9
Chinese Taipei	83.9	63.8	70.5	69.2	80.3
Croatia	75.0	67.5	74.4	68.8	74.9
Cyprus	70.3	63.1	69.2	62.4	74.6
Czech Republic	60.7	56.3	66.6	60.2	69.3
Denmark	62.4	55.5	64.2	49.4	67.0
Finland	73.3	60.5	72.9	62.9	75.7
France	63.0	54.3	63.2	53.2	64.5
Germany	59.2	55.2	66.4	56.7	66.9
Greece	62.8	57.5	61.4	56.2	68.2
Hungary	65.8	55.2	68.2	57.5	70.4
Italy	62.1	53.6	63.1	53.5	62.8
Kazakhstan	65.8	51.4	64.3	58.5	64.9
Korea, Republic of	81.7	59.7	66.4	67.7	74.6
Kosovo	58.9	54.4	63.5	50.6	52.8
Latvia	66.2	52.8	66.5	56.9	68.3
Luxembourg	60.0	54.6	66.3	55.8	65.1
Malta	71.4	58.3	76.2	63.0	69.7
† Netherlands	69.4	57.8	70.3	59.1	68.0
Norway	70.0	60.1	68.0	62.3	73.7
Oman	57.7	50.3	54.5	47.8	49.7
Portugal	68.3	59.9	69.1	66.7	68.0
† Romania	71.3	65.5	71.6	66.3	70.3
Serbia	68.4	62.8	71.5	63.1	72.5
Slovak Republic	64.6	52.8	64.8	63.3	69.1
Slovenia	69.2	66.7	72.4	68.5	74.5
Spain	67.9	56.9	64.8	55.6	67.1
Sweden	75.0	63.3	69.2	62.5	76.1
† United States	65.9	59.6	65.1	53.3	66.7
Uruguay	58.7	56.8	62.6	51.5	61.2
† North Rhine-Westphalia, Germany	60.3	55.7	68.8	59.1	68.8
International Average	66.2	57.3	67.6	60.6	68.4

Note: † These countries are excluded from the pooled international dataset.



Table H.12: Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 1)

Country	T_EMPITE	T_COLICT	T_RESRC	T_VWNEG	T_VWPOS
†Austria	80.1	64.4	66.8	56.8	55.6
Azerbaijan	88.2	73.6	76.7	57.4	64.9
†Belgium (Flemish)	79.3	62.5	67.7	51.4	51.5
†Bosnia and Herzegovina	92.9	81.4	70.4	63.1	60.8
†Chile	83.7	72.8	62.5	53.4	59.9
Chinese Taipei	88.7	78.9	75.6	63.6	68.1
Croatia	93.2	75.3	68.2	59.6	61.5
†Cyprus	87.8	80.2	74.6	61.4	63.1
Czech Republic	86.8	69.9	67.7	54.7	55.5
†Denmark	81.7	59.1	62.9	51.2	55.7
Finland	78.8	66.7	59.9	54.6	54.1
†France	78.8	71.0	63.1	56.3	57.1
†Germany	70.8	56.4	65.2	54.1	53.1
Greece	88.9	81.6	75.5	62.7	64.1
Hungary	91.5	75.9	67.8	59.3	62.8
Italy	81.7	72.7	63.4	63.2	64.0
Kazakhstan	87.7	65.5	74.1	52.3	62.9
Korea, Republic of	81.7	84.6	69.7	63.3	62.2
†Kosovo	70.8	68.2	66.9	50.9	65.7
Latvia	84.1	67.4	60.4	50.5	54.8
†Luxembourg	78.1	61.2	64.7	60.8	59.2
†Malta	78.2	75.5	57.7	55.7	54.9
†Netherlands	72.4	58.6	61.1	51.0	45.0
Norway	81.6	65.9	61.7	48.4	50.4
†Oman	71.4	73.2	69.6	56.0	65.7
Portugal	91.2	69.5	66.7	58.0	66.2
†Romania	72.2	81.8	77.6	60.4	70.6
Serbia	90.9	72.0	70.6	56.1	67.7
†Slovak Republic	87.2	71.3	67.4	54.4	57.0
Slovenia	88.0	69.9	60.8	57.0	53.7
Spain	86.1	73.1	66.8	60.3	64.0
†Sweden	82.4	66.8	59.6	52.2	58.6
†United States	87.8	78.4	72.9	64.2	65.2
†Uruguay	82.8	78.5	68.1	54.3	68.3
†North Rhine-Westphalia, Germany	72.6	60.3	63.4	54.5	54.8
International Average	88.9	73	68	56.5	60.7

† These countries are excluded from the international calibrations. – The items belonging to the scale are not administered in these countries.



Table H.13: Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 2)

Country	T_TEAPTC	T_TEAPIN	T_USETOOL	T_CODEMP	T_EPIEMC
†Austria	56.2	73.1	55.9	67.4	—
Azerbaijan	74.8	72.3	75.5	71.3	65.2
†Belgium (Flemish)	51.0	70.0	47.9	67.4	50.8
†Bosnia and Herzegovina	61.0	79.7	76.6	78.4	61.5
†Chile	59.5	78.8	67.3	76.1	57.6
Chinese Taipei	73.4	76.5	66.2	76.5	79.6
Croatia	64.2	75.5	62.6	71.9	53.5
†Cyprus	65.5	81.4	61.2	75.4	63.2
Czech Republic	57.6	75.2	57.9	68.2	—
†Denmark	49.9	71.6	—	70.4	42.7
Finland	55.2	74.0	43.6	67.7	—
†France	54.7	70.7	53.0	70.3	53.2
†Germany	53.1	71.2	53.1	—	—
Greece	64.9	79.9	64.7	72.5	57.9
Hungary	56.0	78.5	60.8	73.7	57.1
Italy	55.2	73.8	51.6	63.3	55.7
Kazakhstan	58.9	71.3	70.6	70.3	62.8
Korea, Republic of	60.8	76.2	68.8	73.5	75.8
†Kosovo	70.6	75.7	70.6	74.9	63.2
Latvia	51.9	70.1	57.4	71.7	49.0
†Luxembourg	51.2	72.2	50.4	72.7	—
†Malta	53.6	72.6	54.8	71.7	61.0
†Netherlands	53.9	69.3	55.7	70.4	46.5
Norway	49.9	58.7	49.2	66.1	48.8
†Oman	68.2	78.0	72.2	78.6	64.7
Portugal	63.2	75.9	55.7	68.1	56.5
†Romania	74.7	81.5	73.8	79.0	62.4
Serbia	62.4	79.4	66.8	71.3	64.0
†Slovak Republic	62.9	80.8	54.9	72.1	54.4
Slovenia	65.6	76.5	54.9	74.0	61.2
Spain	56.0	76.1	57.3	66.7	59.0
†Sweden	50.5	69.3	50.3	73.5	53.8
†United States	68.8	81.0	55.6	75.5	63.9
†Uruguay	60.4	74.1	60.7	68.9	57.8
†North Rhine-Westphalia, Germany	53.3	70.1	59.0	—	—
International Average	58.1	76.5	64.2	72.4	60.9

† These countries are excluded from the international calibrations. — The items belonging to the scale are not administered in these countries.



Table H.14: Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 3)

Country	T_ICTCLASB	T_ICTCLASA	T_EPICOG	T_ICTEFF	T_PROFLRN
†Austria	63.5	74.3	—	49.1	59.5
Azerbaijan	74.4	77.9	72.0	61.8	78.8
†Belgium (Flemish)	58.6	73.2	54.2	52.4	61.0
†Bosnia and Herzegovina	68.4	83.6	62.2	60.0	74.7
†Chile	66.7	74.7	59.9	62.0	75.3
Chinese Taipei	73.8	80.3	81.5	61.4	65.4
Croatia	64.7	73.7	60.4	63.0	69.1
†Cyprus	68.4	73.4	65.4	64.7	69.2
Czech Republic	64.3	71.3	—	55.2	65.3
†Denmark	55.4	69.7	42.7	47.1	67.7
Finland	57.1	66.0	—	53.9	65.4
†France	52.6	65.9	50.4	51.6	66.4
†Germany	56.8	66.4	—	47.6	55.5
Greece	71.9	78.7	60.5	62.9	68.0
Hungary	70.2	73.3	59.8	54.6	71.1
Italy	60.4	68.1	53.1	55.1	67.9
Kazakhstan	67.4	75.2	71.6	60.7	62.6
Korea, Republic of	72.1	78.7	77.6	64.2	70.1
†Kosovo	69.6	76.0	71.0	56.3	72.7
Latvia	59.2	69.8	59.2	55.5	58.4
†Luxembourg	60.9	65.3	—	49.1	57.2
†Malta	64.2	69.7	61.9	50.1	63.4
†Netherlands	61.9	73.9	49.7	51.9	60.0
Norway	55.1	66.5	48.7	52.5	67.9
†Oman	65.5	79.6	73.0	56.6	66.0
Portugal	67.4	72.9	55.3	60.8	67.1
†Romania	63.8	78.5	65.3	58.3	77.1
Serbia	69.6	79.3	67.5	61.2	72.4
†Slovak Republic	67.7	73.0	57.9	54.0	64.9
Slovenia	63.1	77.3	62.7	52.9	59.3
Spain	64.6	70.1	56.8	59.4	64.0
†Sweden	55.3	66.1	53.7	50.7	62.3
†United States	62.6	71.9	66.0	70.2	69.7
†Uruguay	62.2	71.0	57.2	55.0	66.0
†North Rhine-Westphalia, Germany	53.9	69.4	—	50.0	56.6
International Average	68.5	77.7	64.5	56.9	68.4

† These countries are excluded from the international calibrations. — The items belonging to the scale are not administered in these countries.



Table H.15: Proportion of explained variance by the first principal component by country for scales from teacher questionnaire (part 4)

Country	T_PROFNEED	T_USEUTIL	T_ICTEMP	T_EPICON	T_TEAPEX
†Austria	64.4	51.6	66.9	—	72.3
Azerbaijan	86.6	74.4	74.6	74.8	82.6
†Belgium (Flemish)	63.8	49.6	66.6	60.9	78.6
†Bosnia and Herzegovina	85.7	65.7	78.4	74.2	78.4
†Chile	84.1	59.8	67.5	66.8	76.4
Chinese Taipei	84.9	60.0	72.7	77.6	75.4
Croatia	80.5	56.0	73.0	68.0	77.8
†Cyprus	79.2	54.1	75.5	73.3	81.0
Czech Republic	70.8	50.7	63.7	—	67.0
†Denmark	67.9	—	60.6	48.2	73.4
Finland	67.1	45.1	70.9	—	64.9
†France	67.9	46.1	68.0	56.8	62.8
†Germany	63.1	51.2	61.5	—	72.1
Greece	75.3	62.9	75.6	70.2	80.8
Hungary	76.5	63.0	70.4	61.8	75.6
Italy	74.7	53.3	63.0	58.9	75.3
Kazakhstan	73.5	62.9	64.8	73.8	75.0
Korea, Republic of	80.8	59.5	70.5	74.3	82.0
†Kosovo	82.8	66.4	66.8	72.6	78.8
Latvia	65.8	55.0	66.9	58.6	75.8
†Luxembourg	63.3	48.1	68.9	—	67.5
†Malta	67.7	46.1	66.7	71.4	76.1
†Netherlands	60.4	44.7	67.4	52.9	79.8
Norway	68.4	47.4	61.2	54.4	72.7
†Oman	77.5	61.1	66.2	74.2	82.2
Portugal	73.0	51.0	66.5	59.6	77.1
†Romania	84.0	66.7	73.0	68.5	82.7
Serbia	82.6	57.6	69.1	72.8	82.3
†Slovak Republic	72.8	58.1	66.9	64.7	71.8
Slovenia	72.8	53.4	70.3	69.1	77.7
Spain	70.2	53.6	64.0	61.5	73.7
†Sweden	66.0	47.1	68.5	57.0	71.7
†United States	75.1	52.2	77.1	67.6	78.3
†Uruguay	76.8	57.2	62.5	63.7	70.8
†North Rhine-Westphalia, Germany	63.1	48.8	61.3	—	70.6
International Average	76.4	56.3	69.2	68.0	77.9

† These countries are excluded from the international calibrations. — The items belonging to the scale are not administered in these countries.



Table H.16: Proportion of explained variance by the first principal component for the pooled international dataset for principal and ICT coordinator scales

Scale	Explained variance
C_HINRES	64.1
C_ICTRES	43.1
C_RESTOOL	47.5
C_RESUTIL	70.7
C_HINPED	52.2
P_VWICT	65.1
P_EXPLRN	56.4
P_EXPTCH	78.3
P_PRIORS	60.2
P_ICTUSE	48.3
P_ICTCOM	68.1
P_CIMNEG	51.7
P_CIMPOS	68.9
P_CWORK	54.2
P_CONEG	61.8

Errata note

This note lists the incorrect data spotted in the first edition of the *ICILS 2023 Technical Report*, released on 26 May 2025.

If you require further information about the errata listed here, please contact icils@iea-hamburg.de

Chapter	Table/figure/section	Issue	Correction
1	Table 1.2	A table was added	A table that documents student age and grade level equivalence was added.
2	Table 2.12	A scoring variable was omitted	A scoring variable was inadvertently omitted from Table 2.12. After its inclusion, the number of scoring variables for CIL aspect 4.2 increased from nine to 10. Consequently, the total number of scoring variables across all CIL aspects was updated from 108 to 109. The percentages of scoring variables associated with CIL aspects 3.2 and 4.2 were also revised. Specifically, the percentage for aspect 3.2 was adjusted from 30 percent to 29 percent, and for aspect 4.2, from nine percent to eight percent. The reference to the total number of scoring variables in the description following Table 2.12 was also updated from 108 to 109.
3	Table 2.12	Incorrect description following Table 2.12.	The description following Table 2.12 incorrectly summarized the distribution of score points across CIL strands. The percentage of score points for CIL strand 1 was revised from 13 percent to 12 percent, and the percentage for CIL strand 3 was revised from 48 percent to 49 percent.
4	Section "International parameter calibration" in Chapter 12	Omitted description	The description omitted some specific cases of scales where IRT parameters were derived from international pooled PCM models. The text was revised and more details were added.
5	Tables 12.69, 12.71, 12.73, 12.75, 12.81, 12.83, 12.85, 12.87, 12.89, and 12.95	Missing note below the tables	The note "The IRT parameters were derived from the international pooled PCM model" was added to clarify that the IRT parameters were derived from international pooled PCM models and not multiple group PCM models as this was the case for other scales.



The International Computer and Information Literacy Study (ICILS) 2023 investigated how well students are prepared for life, learning, and work in an increasingly digitalized world. As the third cycle of ICILS, this study has continued to assess international differences in students' computer and information literacy (CIL)—their ability to use digital technologies to investigate, create, communicate, and participate effectively in personal, educational, occupational, and civic contexts. As initiated in ICILS 2018, participating education systems could also choose to assess students' computational thinking (CT), which refers to their capacity to conceptualize real-world problems in computational terms and to develop and evaluate algorithmic solutions that can be executed by digital systems.

This technical report complements the international reports that present the results of ICILS 2023. It provides comprehensive documentation on the conceptual framework and instrument development; translation and verification procedures; sampling design and implementation; the use and management of digital assessment platforms; field operations including data collection quality control; data processing and construction of the international database; and scaling, analysis, and reporting methods.

The report enables researchers to evaluate published findings based on ICILS 2023 data. When used in conjunction with the *ICILS 2023 User Guide for the International Database*, it also serves as a resource for those conducting secondary analyses.

Since its inception in 2013, ICILS has contributed to the global understanding of digital literacy education. The ICILS 2023 database offers rich opportunities for examining how education systems equip young people with essential digital competencies and for evaluating progress toward national and international educational goals, including those articulated in the United Nations' 2030 Agenda for Sustainable Development, and in the European Union's Resolution on a strategic framework for European cooperation in education and training towards the European Education Area and beyond (2021–2030).

