IEA International Computer and Information Literacy Study 2018

TECHNICAL REPORT

Edited by
Julian Fraillon
John Ainley
Wolfram Schulz
Tim Friedman
Daniel Duckworth

IEA International Computer and
Information Literacy Study 2018

Technical Report

Editors

Julian Fraillon
Wolfram Schulz
Daniel Duckworth
John Ainley
Tim Friedman

IEA International Computer and Information Literacy Study 2018

Technical Report

Contributors

John Ainley
Ralph Carstens
Alex Daraganov
Sandra Dohr
David Ebbs
Julian Fraillon
Tim Friedman
Sebastian Meyer
Ekaterina Mikheeva
Lauren Musu
Louise Ockwell
Wolfram Schulz
Sabine Tieck

Julian Fraillon
The Australian Council for Educational Research
Camberwell, Victoria
Australia

John Ainley
The Australian Council for Educational Research
Camberwell, Victoria
Australia

Wolfram Schulz
The Australian Council for Educational Research
Camberwell, Victoria
Australia

Tim Friedman
The Australian Council for Educational Research
Camberwell, Victoria
Australia

Daniel Duckworth
The Australian Council for Educational Research
Camberwell, Victoria
Australia

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

ISBN/EAN: 9789079549351

IEA
Researching education, improving learning

The International Association for the Evaluation of Educational Achievement (IEA), with headquarters in Amsterdam, 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.

Design by Becky Bliss Design and Production, Wellington, New Zealand
Cover design by Studio Lakmoes, Arnhem, The Netherlands

Contents

List of tables and figures

Tables

Figures

CHAPTER 1:

Overview of the IEA International Computer and Information Literacy Study 2018

Julian Fraillon, Sebastian Meyer, and John Ainley

Introduction

The International Association for the Evaluation of Educational Achievement (IEA) International Computer and Information Literacy Study (ICILS) 2018 investigated how well students are prepared for study, work, and life in a digital world. There is increasing acknowledgment across countries that with rapid advancement of new technologies it is important to develop the capacities of people to use information and communication technologies (ICT) (see, for example, European Commission 2018). ICILS 2018 focused on "the capacities of students to use ICT productively for a range of purposes, in ways that go beyond a basic use of ICT" (Fraillon et al. 2019, p. 1). ICILS 2018 was based on and expanded the work of ICILS 2013 (Fraillon et al. 2014).

ICILS 2018 included three main focus areas. Firstly, ICILS 2018 assessed computer and information literacy (CIL) which was 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 et al. 2019, p. 16). CIL was also assessed in the first study cycle, ICILS 2013. Secondly, as an optional component for participating countries, ICILS 2018 assessed computational thinking (CT) which 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" (Fraillon et al. 2019, p. 27). CT was not assessed in ICILS 2013. The assessment of CT was an innovative and engaging challenge for the students, evaluating not only their ability to analyze and break down a problem into logical steps but also assessing their understanding of how computers might be used to solve a problem. Thirdly, ICILS 2018 investigated the contexts in which students' CIL and CT are developed by collecting and analyzing data relating to the use of computers and other digital devices by students and teachers, and the resources available to support the teaching and learning of CIL and CT in schools.

ICILS was based around four research questions concerned with: (1) variations in CIL and CT between and within countries; (2) aspects of schools, education systems, and teaching that are related to student achievement in CIL and CT; (3) the extent to which students' access to, familiarity with, and self-reported proficiency in using ICT are related to student achievement in CIL and CT; and (4) the aspects of students' personal and social backgrounds that are related to CIL and CT.

The ICILS 2018 assessment framework (Fraillon et al. 2019) describes the development of these research questions. The framework also provides greater detail relating to the measured domains and outlines the variables necessary for analyses associated with the research questions.

ICILS systematically reviewed differences among participating countries and education systems[1] with regard to students' CIL and CT, and how participating countries and systems provided and supported ICT-related education. It explored differences within and across countries with respect to the relationship between ICT-related learning outcomes, student characteristics, and school contexts. The outcomes of these reviews and analyses are reported in the ICILS 2018 international report (Fraillon et al. 2020).

1 Education systems are units within countries with a degree of educational autonomy that have participated following the same standards for sampling and testing as countries. In this report, education systems are often referred to as countries for ease of reading.

Instruments

CIL test

Students completed a computer-based test of CIL that consisted of questions and tasks that were administered as five different 30-minute modules. Three of the CIL modules had been developed and used in ICILS 2013 and kept secure as *trend modules*. These were included to allow data collected in ICILS 2018 to be equated with data from the previous cycle and reported on the CIL proficiency scale established for ICILS 2013. Consequently, it was possible to compare CIL achievement over time in those countries that participated in both cycles. Two new modules were developed for the ICILS 2018 CIL test instrument to address contemporary thematic content and software environments. Data collected from all five CIL modules in ICILS 2018 were used as the basis for reporting ICILS 2018 CIL results on the ICILS CIL achievement scale established in 2013.

Each student completed two modules randomly allocated from the set of five available modules so that the total assessment time for each student was one hour. Each of the assessment modules consisted of a set of questions and tasks based on a realistic theme and following a linear narrative structure. These modules consisted of a series of small discrete tasks (typically taking less than a minute to complete) followed by a large task that typically took 15 to 20 minutes to complete. In total, the modules comprised 81 discrete questions that generated 102 score points.

When students began each module they were first presented with an overview of the theme and purpose of the tasks in the module including a basic description of what the large task would comprise. In the narrative of each module the smaller discrete tasks typically comprised a mix of skill execution and information management tasks that built towards completion of the large task. Students were required to complete the tasks in the allocated sequence and could not return to review completed tasks.

The five modules measuring students' CIL were:

- *Band competition* (2013 & 2018): Students planned a website, edited an image, and used a simple website builder to create a webpage with information about a school band competition.
- *Breathing* (2013 & 2018): Students managed files, and evaluated and collected information in order to create a presentation explaining the process of breathing to eight- or nine-year-old students.
- *School trip* (2013 & 2018): Using online database tools, students helped plan a school trip and selected and adapted information to produce an information sheet about the trip for their peers. The information sheet included a map created using an online mapping tool.
- *Board games* (2018): Students use a school-based social network for direct messaging and group posting to encourage peers to join a board games interest group.
- *Recycling* (2018): Students access and evaluate information from a video sharing website to identify a suitable information source relating to waste reduction, reuse, and recycling. Students take research notes from the video and use their notes as the basis for designing an infographic to raise awareness about waste reduction, reuse, and recycling.

In total, there were 20 different possible combinations of module pairs. Each module appeared in eight of the combinations—four times as the first and four times as the second module when paired with each of the other four. The module combinations were randomly allocated to students. This test design made it possible to assess a larger amount of content than could be completed by any individual student and thus ensured broad coverage of the content of the ICILS 2018 assessment framework. The design also controlled for item position effects on task difficulty across the sampled students and provided a variety of contexts for the assessment of CIL.

CT test

Students in those countries participating in the CT assessment completed an additional computer-based test of CT that consisted of two 25-minute modules. These modules were developed for ICILS 2018 and corresponded to the two strands of the CT construct, one of which was related to conceptualizing problems and the other to operationalizing solutions. Each module had a unifying theme and a sequence of related tasks. Unlike the CIL modules, each CT module did not culminate in a large task. Students completed the two CT test modules in a randomized order after they had completed both the CIL test and the student questionnaire. In total, the CT data comprised 39 score points derived from 18 discrete tasks and questions. Student responses to most tasks were automatically scored. The exceptions were some open-response questions that were scored by trained scorers at each national center.

The tasks in the CT module primarily focused on planning digital solutions needed to control a driverless bus. The set of tasks included manipulating and interpreting visual representations of information and processes associated with behavior of the bus, and configuring simulations to collect data and draw conclusions about the behavior of the bus under specified conditions.

In the CT module focusing on operationalizing solutions, students worked within a simple visual coding environment comprising blocks of code that have some specified and some configurable functions and could be assembled in sequence to control the actions of a simulated drone used in farming. Students were required to create, test, and debug code-based algorithms. Students were presented with a work space, "draggable" commands, and a visual output that showed the drone completing the commands. The complexity of each task related to the number of targets and actions required to solve the problem instance. The tasks were more complex as the students advanced through the module. The complexity of the tasks related to the variety of code functions that were available and the sequence of actions required by the drone for completion of the task.

Questionnaires

The completion of the CIL assessment was followed by a 30-minute student questionnaire also administered on computer. The questionnaire included questions relating to students' background characteristics, their experience and use of ICT to complete a range of different tasks in school and out of school, and their attitudes towards the use of ICT.

Three further instruments were designed to gather information from and about teachers and schools:

- A *30-minute teacher questionnaire* included some questions relating to teachers' background followed by questions relating to teachers' reported familiarity with ICT, their use of ICT in educational activities in teaching focused on a "reference class,"[2] their perceptions of ICT in schools, and their participation in professional learning activities relating to their use of ICT when teaching.
- A *15-minute ICT coordinator questionnaire* asked ICT coordinators about the resources available in their school to support the use of ICT in teaching and learning. In addition, the questionnaire gathered information about schools' technological (e.g., infrastructure, hardware, and software) and pedagogical support as well as professional learning and hindrances to the use of ICT in school education.
- A *15-minute principal questionnaire* asked principals to provide information about school characteristics and then about school approaches to ICT-related teaching and incorporating ICT into teaching and learning.

2 The "reference class" was defined as the first target grade class taught by the respondent for a regular subject (i.e., other than home room, assembly, etc.) on or after the Tuesday following the last weekend before the respondent first accessed the questionnaire. See page 7 for a complete definition of the reference class.

An additional *national context questionnaire* was used to gather information from ICILS 2018 national centers about national contexts for and approaches to the development of students' CIL and CT. This included information on policies and practices as well as on expectations and requirements for the pedagogical use of ICT. When answering this questionnaire, which was administered online, national centers were requested to draw on all available national expertise to provide the required information as well as to provide reference documents where appropriate.

Computer-based test delivery

ICILS 2018 used purpose-designed software for the computer-based student assessment and questionnaire. These were administered primarily using USB drives connected to school computers. After administration of the student instruments, the ICILS research team either directly uploaded data to a server or submitted this information to national research centers for subsequent upload by national center staff.

The teacher and school questionnaires were usually completed online (directly accessing a server at IEA over the internet). However, respondents were also offered the option of completing the questionnaires on paper.

Measures

The CIL scale

CIL was 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 et al. 2019, p. 16). The ICILS 2018 CIL construct was conceptualized around four strands that framed skills and knowledge addressed by the CIL instruments.[3] We used the Rasch item response theory model (Rasch 1960) to derive the CIL scale from the data collected from student responses to 81 test questions and large tasks that generated 102 score points. Most questions and tasks each corresponded to one item. However, raters scored each ICILS large task against a set of criteria (each criterion with its own unique set of scores) relating to the specific properties of the task. Each large task assessment criterion can therefore be regarded as an item in ICILS.

The ICILS CIL reporting scale was established in ICILS 2013, with a mean of 500 (the average CIL scale score across countries in 2013) and a standard deviation of 100 for the equally weighted national samples that met IEA sample participation standards in the first cycle. Plausible values were generated with full conditioning to derive summary student achievement statistics.

The described scale of CIL achievement in ICILS is based on the content and scaled difficulties of the assessment items. The ICILS research team wrote descriptors for the expected CIL knowledge, skills, and understandings demonstrated by students who correctly responded to these items. Ordering the item descriptors according to their scaled difficulty (from least to most difficult) was used to develop an item map. The content of the items was used to inform judgements about the skills represented by groups of items on the scale ordered by difficulties.

Analysis of this item map and the student achievement data were then used to establish proficiency levels that each had a width of 85 scale points with level boundaries set at 407, 492, 576, and 661 scale points (rounded to the nearest whole number). Student scores below 407 scale points indicate CIL proficiency below the lowest level targeted by the assessment instrument.

3 CIL was described as comprising only two strands in ICILS 2013. Following an extensive evaluation of the ICILS 2013 CIL construct and in consultation with ICILS national researchers the ICILS project team established a revised structure for the CIL construct for ICILS 2018. The restructuring of the CIL construct was undertaken to better communicate the contents and emphases of the construct and to minimize overlap across the aspects of the construct (Fraillon et al. 2019, p. 17).

The four described levels of the CIL scale are summarized as follows:

- *Level 4* (above 661 scale points): Students working at Level 4 select the most relevant information to use for communicative purposes. They evaluate usefulness of information based on criteria associated with need and evaluate the reliability of information based on its content and probable origin. These students create information products that demonstrate a consideration of audience and communicative purpose. They also use appropriate software features to restructure and present information in a manner that is consistent with presentation conventions. They then adapt that information to suit the needs of an audience. Students working at Level 4 demonstrate awareness of problems that can arise regarding the use of proprietary information on the internet.
- *Level 3* (577 to 661 scale points): Students working at Level 3 demonstrate the capacity to work independently when using computers as information gathering and management tools. These students select the most appropriate information source to meet a specified purpose, retrieve information from given electronic sources to answer concrete questions, and follow instructions to use conventionally recognized software commands to edit, add content to, and reformat information products. They recognize that the credibility of web-based information can be influenced by the identity, expertise, and motives of the creators of the information.
- *Level 2* (492 to 576 score points): Students working at Level 2 use computers to complete basic and explicit information gathering and management tasks. They locate explicit information from within given electronic sources. These students make basic edits and add content to existing information products in response to specific instructions. They create simple information products that show consistency of design and adherence to layout conventions. Students working at Level 2 demonstrate awareness of mechanisms for protecting personal information and some consequences of public access to personal information.
- *Level 1* (407 to 491 score points): Students working at Level 1 demonstrate a functional working knowledge of computers as tools and a basic understanding of the consequences of computers being accessed by multiple users. They apply conventional software commands to perform basic research and communication tasks and add simple content to information products. They demonstrate familiarity with the basic layout conventions of electronic documents.

The CT scale

CT refers to an "individual's 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" (Fraillon et al. 2019, p. 27). The CT construct comprised two strands: conceptualizing problems and operationalizing solutions.

We used the Rasch item response theory model (Rasch 1960) to derive the CT scale from student responses to 18 CT tasks that generated 39 score points. The final reporting CT scale was set to a metric that had a mean of 500 (the ICILS average score) and a standard deviation of 100 for the equally weighted national samples in countries who administered the CT modules. As for CIL, we used plausible values with full conditioning to derive summary student achievement statistics for CT.

The ICILS described scale of CT achievement is based on the content and scaled difficulties of the assessment items. The ICILS research team wrote descriptors for the expected CT knowledge, skills, and understandings demonstrated by students correctly responding to each item. Ordering the item descriptors according to their scaled difficulty (from least to most difficult) resulted in an item map, similar to the item map that was produced for CIL.

Given the limited number of CT tasks and score points, it was not possible to establish proficiency levels in the same way as for CIL. However, to provide a broad description of the underlying characteristics of achievement across the breadth of the scale we divided the items ordered by their difficulty into thirds with equal numbers of items in each third. For ICILS 2018 we refer to these as the lower, middle, and upper regions of the CT scale. The descriptions of each region are syntheses of the common elements of CT knowledge, skills, and understanding described by the items within each region. The regions of the CT scale cannot be directly compared to the levels in the CIL scale, as they have been developed using a different process and the scale metrics are not comparable.

The three regions of the CT scale can be described as follows:

- *Upper region* (above 589 scale points): Students showing achievement corresponding to the upper region of the scale demonstrate an understanding of computation as a generalizable problem-solving framework. They can explain how they have executed a systematic approach when using computation to solve real-world problems. Furthermore, students operating within the upper region can develop algorithms that use repeat statements together with conditional statements effectively.
- *Middle region* (459 to 589 scale points): Students showing achievement corresponding to the middle region of the scale demonstrate understanding of how computation can be used to solve real-world problems. They can plan and execute systematic interactions with a system so that they can interpret the output or behavior of the system. When developing algorithms, they use repeat statements effectively.
- *Lower region* (below 459 scale points): Students showing achievement corresponding to the lower region of the scale demonstrate familiarity with the basic conventions of digital systems to configure inputs, observe events, and record observations when planning computational solutions to given problems. When developing problem solutions in the form of algorithms, they can use a linear (step-by-step) sequence of instructions to meet task objectives.

Measures based on the student questionnaire

A number of the measures based on the student questionnaire were single-item indices based on student responses:

- Experience with using computers' (made up of five categories based on years of using computers);
- Frequency of computer use at home, school, and other places (made up of five categories);
- Frequency of use of various applications (made up of five categories); and
- Frequency of use of computers in different subject areas (four categories and eight subject areas).

In addition, students responded to sets of items that had been designed to measure constructs of interest and provided the basis for generating scales reflecting these underlying latent traits. Scales were developed to provide measures of a number of dimensions concerned with ICT engagement. The scales included measures of the extent of use of ICT for various purposes and of student perceptions of ICT. Measures of the extent of use of ICT included the use of ICT for general applications, school purposes, communication and information exchange, and leisure. Measures of perceptions included ICT self-efficacy (in relation to general and specialist ICT applications) and perceptions of effects of ICT on society. Scales were also generated to reflect the extent to which students learned about CIL and CT tasks at school.

Measures based on the teacher questionnaire

A number of the measures based on the teacher questionnaire were also single-item indices. Such measures included experience with using computers for teaching purposes, and frequency of computer use in and outside school for pedagogical and other purposes.

Sets of items were designed to generate scales reflecting underlying constructs such as:

- Teachers' perceptions of school resources for ICT;
- Teachers' views of ICT for teaching and learning;
- Teachers' self-confidence in using ICT (for general and specialist applications); and
- Teachers' collaboration with other teachers about how ICT is used.

In addition, in order to determine measures of the extent to which the teachers were using ICT in their teaching, the teacher questionnaire asked the teachers what they did in a particular reference class. Teachers were asked to select the reference class from among the classes each of them was teaching, and to base their responses regarding their teaching practices on their experiences with that particular class. To ensure that the selection was unbiased, the following instruction was provided:

> This is the first [target grade] class that you teach for a regular subject (i.e., other than home room, assembly etc.) on or after Tuesday following the last weekend before you first accessed this questionnaire. You may, of course, teach the class at other times during the week as well. If you did not teach a [target grade] class on that Tuesday, please use the [target grade] class that you taught on the first day after that Tuesday.

Teachers provided information on how frequently they used ICT in the reference class, the subject area they taught to the reference class, the emphasis they placed on developing students' CIL, the ICT tools that they used, the learning activities in which they used ICT, and the teaching practices in which they incorporated ICT.

Measures based on the school questionnaire

The school questionnaires (for the school principal and the ICT coordinator) provided measures of school access to ICT resources, school policies and practices for using ICT, impediments to using ICT in teaching and learning, and participation in teacher professional development. In addition, those questionnaires provided information about school characteristics, school contexts, and ICT resources.

Data from school questionnaires provided information that was used to derive both simple indices (for example, on ratios of school size and the number of ICT devices) and scales.

Data

Countries

The twelve countries that participated in ICILS were: Chile, Kazakhstan, Denmark, the Republic of Korea (hereafter referred to as Korea, for ease of reading), Finland, Luxembourg, France, Portugal, Germany, the United States, Italy, and Uruguay. Moscow (Russian Federation) and North Rhine-Westphalia (Germany) took part as benchmarking participants.

Denmark, Korea, Finland, Luxembourg, France, Portugal, Germany, the United States, and North Rhine-Westphalia (Germany) participated in the CT international option.

Population definitions

The ICILS student population was defined as students in grade 8 (typically around 14 years of age in most countries), provided that the average age of students in this grade was at least 13.5 at the time of the assessment.

The population for the ICILS teacher survey was defined as all teachers teaching regular school subjects to the students in the target grade. It included only those teachers who were teaching the target grade during the testing period and who had been employed at school since the beginning of the school year. ICILS also administered separate questionnaires to principals and nominated ICT coordinators in each school.

Sample design

The samples were designed as two-stage cluster samples. During the first stage of sampling, PPS procedures (probability proportional to size, as measured by the number of students enrolled in a school) were used to sample schools within each country. The numbers required in the sample to achieve the necessary precision were estimated on the basis of national characteristics. However, as a guide, each country was instructed to plan for a minimum sample size of 150 schools.[4] Sample sizes of schools within countries ranged between 143 and 210 across countries. The sampling of schools constituted the first stage of sampling both students and teachers.

Twenty students were then randomly sampled from all students enrolled in the target grade in each sampled school. In schools with fewer than 20 students, all students were invited to participate.

All teachers of the target grade were eligible to be sampled regardless of the subjects they taught given that ICT use for teaching and learning is not restricted to particular subjects. In most schools (those with 21 or more teachers of the target grade), 15 teachers were selected at random from all teachers teaching the target grade. In schools with 20 or fewer such teachers, all teachers were invited to participate.

The sample participation requirements were applied independently to students and teachers in schools. The requirement was 85 percent of the selected schools and 85 percent among the selected participants (students or teachers) within the participating schools, or a weighted overall participation rate of 75 percent.

Achieved samples

ICILS gathered data from more than 46,000 lower-secondary students in more than 2200 schools from 14 countries or education systems within countries. These student data were augmented by data from more than 26,000 teachers in those schools and by contextual data collected from school ICT coordinators, school principals, and national research centers. Eleven out of 12 countries and both benchmarking participants met IEA sample participation standards for the student survey, for the teacher survey this was the case for seven out of 12 countries and both benchmarking participants. Data from countries that did not meet IEA sample participation requirements were reported in separate sections of the reporting tables.

The main ICILS survey took place in the 14 participating countries and benchmarking participants between February and December 2018. The survey was carried out in countries with a Northern Hemisphere school calendar between February and June 2018 and in those with a Southern Hemisphere school calendar between October and December 2018.[5] In a few cases, ICILS granted countries an extension to their data-collection periods or collected data from their target grade at the beginning of the next school year.

4 In Luxembourg the total amount of eligible schools was below 50 which is why all schools were selected (census).

5 Italy decided to survey students and their teachers at the beginning of the school year while all other countries administered the survey at the end of school year. Their results were annotated accordingly in the international report.

Outline of the technical report

This overview of ICILS 2018 is followed by 13 chapters. Chapters 2, 3, and 4, cover the instruments that were used in the study. Chapter 2 focuses on the development of the tests while Chapter 3 provides an account of the computer-based assessment systems. 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. An appreciation of the material in these chapters provides an essential foundation for interpreting the results of the study.

Chapters 5 through 9 focus on the implementation of the survey in 2018. Chapter 5 describes the translation procedures and national adaptations used in ICILS. Chapter 6 details the sampling design and implementation, while Chapter 7 describes the sampling weights that were applied and documents the participation rates that were achieved. Chapter 8, which describes the field operations, is closely linked to Chapter 9, which reports on the feedback and observations gathered from the participating countries during the data collection.

Chapters 10, 11, and 12 are concerned with data management and analysis. Chapter 10 describes the data-management processes that resulted in the creation of the ICILS database. Chapter 11 details the scaling procedures for the CIL test, i.e., how the responses to tasks and items were used to generate the scale scores and proficiency levels. Chapter 12 describes the scaling procedures for the questionnaire items (student, teacher, and school questionnaires). The final chapter, Chapter 13, presents an account of the analyses that underpinned the international report of ICILS 2018 (Fraillon et al. 2020).

References

European Commission. (2018). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions* (COM(2018) 22 final). Brussels, Belgium: Author. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0022&from=EN

European Council. (2017). *European Council Conclusions, 19 October 2017* (EUCO 14/17). Brussels, Belgium: Author. Retrieved from https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf

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*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783319142210

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030193881

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*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030387808

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

CHAPTER 2:

ICILS 2018 test development

Julian Fraillon

Introduction

The new content for inclusion in the ICILS 2018 assessment was developed over a 20-month period from April 2015 to December 2016. Most of this work was conducted by the international study center (ISC) at ACER in collaboration with national research coordinators (NRCs) and other research partners.

The ICILS 2018 assessment included two tests. The test of computer and information literacy (CIL) was completed by all students and the test of computational thinking (CT) was completed by all students in countries that elected to undertake this additional assessment (see Chapter 1 for further details of country participation in ICILS).

This chapter provides a detailed description of the test development process and review procedures as well as the test design implemented for the ICILS 2018 field trial and main survey. The processes are relevant for the tests of both CIL and CT and are consequently described together. Where relevant, details of the content of each test are described separately.

Table 2.1 provides an overview of the test development processes and timeline.

Test scope and format

ICILS 2018 assessment framework

The ICILS student tests of CIL and CT were developed with reference to the ICILS 2018 assessment framework[1] (Fraillon et al. 2019). CIL was 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 et al. 2019, p. 16) and CT was 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" (Fraillon et al. 2019, p. 27).

Each of CIL and CT are described in the ICILS 2018 assessment framework in terms of *strands* (overarching conceptual categories) and *aspects* (specific content categories within strands). However, the described structures of CIL and CT were not intended to presuppose a sub-dimensional analytic structure (see Fraillon et al. 2019 for further details).

The CIL framework

The following list sets out the four strands and corresponding aspects of the CIL framework. Full details of the CIL construct can be found in the ICILS 2018 assessment framework.

- 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

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

- Strand 3: Producing information, comprising two aspects:
  - 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 framework

The following list sets out the two strands and corresponding aspects of the CT framework. Full details of the CT construct can be found in the ICILS 2018 assessment framework.

- 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 questions and tasks making up the ICILS CIL test instrument were presented in five modules, each of which took 30 minutes to complete. Each student completed two modules randomly allocated from the set of five. Three of the modules were secure trend modules first used in ICILS 2013 that provided a basis for reporting all CIL data collected in ICILS 2018 on the ICILS CIL achievement scale that was established in 2013. Two of the modules were newly developed for inclusion in ICILS 2018.

A module is a set of tasks based on an authentic theme and following a linear narrative structure. Each module has a series of discrete tasks, each of which typically takes less than a minute to complete, followed by a large task that typically takes 15 to 20 minutes to complete. The narrative of each module positions the discrete tasks as a mix of skill-execution and information-management tasks that students need to do in preparation for completing the large task.

When beginning each module, students were presented with an overview of the theme and purpose of the tasks in the module as well as a basic description of what the large task would comprise. Students were required to complete the tasks in the allocated sequence and could not return to review completed tasks. Table 2.2 includes a summary of the five ICILS CIL assessment modules including the large tasks.

The ICILS CIL test modules included three broad categories of task 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" (Fraillon et al. 2019, 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" (Fraillon et al. 2019, p. 46).

Table 2.1: Test development processes and timeline

Year	Month	Group	Activity
2015	February	ICILS international study center	Establishment of CIL test specifications and preliminary CT test specifications
2015	March	First meeting of national research coordinators (Krakow)	Reflections on ICILS 2013 test of CIL Review of proposed test development process and test specifications Test development workshop
2015	March	ICCS international study center	Drafting, review, and refinement of test modules
2015	December	National research coordinators	Web-based review of test module storyboards (1)
2016	January	ICILS international study center	Revision of test module storyboards
2016	February	Second meeting of national research coordinators (Amsterdam)	Review of draft test modules
2016	March	ICILS international study center	Revision of draft test modules following second meeting of national research coordinators and pilot-testing of selected module content
2016	April	National research coordinators	Web-based review of test module storyboards (2)
2016	May	ICILS international study center IEA	Revision of test modules and development of content in online test delivery system
2016	September	Third meeting of national research coordinators (Porto)	Review of modules proposed for inclusion in field trial test and confirmation of test design
2016	October	ICILS international study center	Finalization of field trial test modules
2016	November	ICILS field trial scoring trainers	Review of field trial scoring guides for constructed-response items and large tasks (as part of scoring training)
2016	December	ICILS international study center	Revision of field trial scoring guides for constructed-response items and large tasks
2017	July	ICILS international study center RM Results	Migration of test content to online delivery system
2017	August	ICILS international study center	Analysis of field trial item data and recommendations for modules/items to be included in main survey test (field trial analysis report)
2017	September	Fourth meeting of national) research coordinators (Berlin)	Review of field trial analysis report and recommendations for test design and modules/items proposed for inclusion in main survey
2017	December	ICILS main survey scoring trainers (Hamburg)	Review of main survey scoring guides for constructed-response items and large tasks (as part of scoring training)
2018	January	ICILS international study center	Finalization of main survey scoring guides for constructed-response items and large tasks

- Skills tasks: In these tasks students "use interactive simulations of generic software or universal applications to complete an action" (Fraillon et al. 2019, p. 47). 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 online database to locate information).
- Authoring tasks: These tasks "require students to modify and create information products using authentic computer software applications" (Fraillon et al. 2019, p. 49). The complexity of authoring tasks vary according to the number of different applications students were required to use, the range of viable solutions to the task, and the amount of information students were required to evaluate and make use of when completing the task. The authoring tasks were most commonly the large task within each module and typically were scored using multiple analytic criteria applied by human scorers.

Further details of the ICILS CIL test modules, including example tasks, are presented in the ICILS 2018 assessment framework (Fraillon et al. 2019) and the ICILS 2018 international report (Fraillon et al. 2020).

Table 2.2: Summary of ICILS 2018 CIL test modules and large taskst

Module	Description and large task
Band competition (also used in ICILS 2013)	Students plan a website, edit an image, and use a simple website builder to create a webpage with information about a school band competition.
Breathing (also used in ICILS 2013)	Students manage files and evaluate and collect information to create a presentation to explain the process of breathing to eight- or nine-year-old students.
School trip (also used in ICILS 2013)	Students help plan a school trip using online database tools and select and adapt information to produce an information sheet about the trip for their peers. The information sheet includes a map created using an online mapping tool.
Board games (new for ICILS 2018)	Students use a school-based social network for direct messaging and group posting to encourage peers to join a board games interest group.
Recycling (new for ICILS 2018)	Students access and evaluate information from a video sharing website to identify a suitable information source relating to waste reduction, reuse, and recycling. Students take research notes from the video and use their notes as the basis for designing an infographic to raise awareness about waste reduction, reuse, and recycling.

The CT test instrument

The tasks making up the ICILS CT test instrument were presented in two modules, each of which took 25 minutes to complete. Each student completed both modules (in randomized order across students) after they had completed each of the CIL test and the student questionnaire. Both modules were newly developed for inclusion in ICILS 2018.

Similar to the CIL assessment, each CT test module contained a set of tasks linked by a common narrative theme. Unlike the CIL modules, the CT modules did not culminate in students completing a large task, in contrast, each module comprised a set of items associated with the processes of planning and execution of computer-based solutions to real-world problems.

One of the modules, *automated bus*, focused on content associated with the *conceptualizing problems* strand of the CT framework. The narrative theme of the module related to planning aspects of programmable decision-making to be implemented in a driverless bus, such as route planning and braking at safe distances to avoid collisions. The tasks in this module included visual representation of real-world scenarios (through, for example, path diagrams, flow charts, and decision trees) and configuring, running, and interpreting results of a simulation tool.

The second module, *farm drone*, focused on content associated with the *operationalizing solutions* strand of the CT framework. The narrative theme of this module related to the use of visual code to control the actions of a drone used in farming. In the farm drone module students were able to return to earlier tasks to check and revise their responses.

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 CT assessment. 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 outcomes of the application of algorithmic instructions can be displayed visually" (Fraillon et al. 2019, p. 51). 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 work with some form of simulation tool, typically as part of developing an understanding of real-world problems or to evaluate solutions to problems. The tasks can have students set parameters on the tool, run a simulation, collect data, and interpret the results.
- *Visual coding tasks:* These tasks require students to manipulate visual code blocks that can be used to execute a range of actions. In ICILS 2018, these tasks focused on managing code blocks that could control the movement and some basic actions of a virtual drone. Students could assemble the code blocks in a work space, rearrange the code blocks, and configure some aspects of the code blocks (such as the number of repeats in a loop or the material dropped by drone). Students could also run the code at any time and view the activation of code blocks and the corresponding behavior of the drone. They could also separately reset both the drone and the code blocks/work space to their original states. There were two main forms of visual coding tasks:
  i. *Algorithm construction tasks* require students to assemble sequences of code blocks in order for the drone to execute a prescribed set of actions.
  ii. *Algorithm debugging tasks* require students to correct an existing flawed algorithm (provided to students as an editable configuration of code blocks in the work space) so that the drone could execute a prescribed set of actions.

Further details of the ICILS CT test modules, including example tasks, are presented in the ICILS 2018 assessment framework (Fraillon et al. 2019) and the ICILS 2018 international report (Fraillon et al. 2020).

Test-development process

The test-development process consisted of a series of stages applicable to the development of both the CIL and CT test instruments. Although these stages followed each other sequentially, the iterative and collaborative nature of the overall process meant that some materials were reviewed and revised within particular stages more than once. In summary, the ISC developed item materials (sometimes based on suggestions from NRCs), which were reviewed by NRCs and then revised by the ISC. Sometimes this process was repeated.

In ICILS, each task or item comprises the stimulus materials available to students, the question or instructions given to students, the specified behavior of the computer delivery system in response to students' actions, and the scoring logic (as specified in the scoring guides for human scoring) for each item or task. The test development and review process encompassed all of these constituent parts of the ICILS modules.

Drafting of preliminary module ideas

The first meeting of the ICILS 2018 NRCs, included a reflection on the CIL test from ICILS 2013, discussion of the potential for assessing CT in ICILS 2018, and a module development workshop in which participants were introduced to the questions and issues directing the creation and evaluation of the modules and tasks. These review criteria, which remained valid throughout the module development and review process, were applied by test development staff at the ISC and reviewers alike. The following lists present the main review questions used to evaluate the ICILS modules:

- *Content validity*
  - How did the material relate to the ICILS test specifications?
  - Did the tasks test the construct (CIL or CT) described in the assessment framework?
  - Did the tasks relate to content at the core of the aspects of the assessment framework or focus on trivial side issues?
  - How would the ICILS test content stand up to broader expert and public scrutiny?
- *Clarity and context*
  - Were the tasks and stimulus material coherent, unambiguous, and clear?
  - Were the modules and tasks interesting, worthwhile, and relevant?
  - Did the tasks assume prior knowledge and, if so, was this assumed to be acceptable or part of what the test intended to measure?
  - Was the reading load as low as possible without compromising the real-world relevance and validity of the tasks?
  - Were there idioms or syntactic structures that may prove difficult to translate into other languages?
- *Test-takers*
  - Did the content of the modules and tasks match the expected range of ability levels, age, and maturity of the ICILS target population?
  - Did the material appear to be cross-culturally relevant and sensitive?
  - Were 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?
  - Did the constructed-response items and the large-task information provide clear guidance about what was expected in response to the items and tasks?
- *Format and scoring*
  - Was the proposed format the most suitable for the framework content being assessed by each task?
  - Was the key (the correct answer to a multiple-choice question) indisputably correct?
  - Were the distractors (the incorrect options to a multiple-choice question) plausible but also irrefutably incorrect?
  - Did the scoring criteria for the large tasks assess the essential characteristics of task completion?
  - Were there different approaches to provide answers with the same score, and did they represent equivalent or different levels of proficiency?
  - Was the proposed scoring consistent with the underlying ability measured by the test (CIL), and would test respondents with higher ability levels always score better than those with lower ones?
  - Were there other kinds of answers that had not been anticipated in the scoring guides (e.g., any that did not fall within the "correct" answer category description, but appear to be equally correct)?

– Were the scoring criteria sufficient for scorers and did they clearly distinguish the different levels of performance?

At the module development workshop, NRCs were invited to discuss and suggest new ideas for module themes. These themes were taken as a starting point for subsequent module development.

Development of module storyboards

After the first NRC meeting, the ISC developed storyboards for six new modules (three for each of CIL and CT).

The storyboards were presented in the form of a Microsoft PowerPoint mock-up of each task/item 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 PowerPoint slide contained the stimulus material together with the task instructions or question that students would be required to respond to, and each storyboard was accompanied by a set of implementation notes for each task or question. The notes 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.

First online review of module storyboards

In December 2015, NRCs took part in an online review of the draft storyboards. When reviewing the draft storyboards, NRCs made recommendations relating to the modules. The NRCs' feedback informed further revision of the module storyboards and preparation for detailed discussion of the modules at the second meeting of NRCs.

Face-to-face review of draft module storyboards

One focus of the second meeting of NRCs was to review the six draft module storyboards.

Following this meeting, four module storyboards (two CIL and two CT modules) were selected for further development and revision. Feedback from the meeting was used to inform the further revision of the four module storyboards.

Second online review of draft module storyboards

Once the module storyboards had been revised following the second meeting of NRCs, the storyboards were made available online to NRCs for further comment. The feedback on these modules was used to finalize the draft storyboards to be enacted within the online delivery system.

Authoring the draft storyboards in the field trial online delivery system

The finalized storyboards were provided to IEA to be authored into the test delivery system, which meant they could then be viewed, in draft form, with their expected functionality. This process served two purposes: it enabled the draft modules to be reviewed and refined with reference to their live functionality and it enabled the functionality of the test delivery system to be tested and refined.

Face-to-face review of field-trial modules and finalization

Operational versions of the proposed field trial modules were made available to NRCs for review at the third meeting of NRCs. The content and functionality of the field trial modules were revised and finalized in response to this review.

Field-trial scorer training

National center representatives attended an international scorer training meeting 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 scoring training process led to refinements to the scoring guides.

Authoring the draft storyboards in the main survey online delivery system

Following the field trial, the module content was provided to RM Results (formerly SoNET Systems) 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.

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

Field trial data were used to investigate the measurement properties of the ICILS test items at the international level and within countries. Having recommended which modules and tasks should be included in the main survey instrument, ISC staff discussed their recommendations with NRCs at the fourth ICILS NRC meeting. During the meeting, small refinements were recommended for a number of tasks. The NRCs also strongly recommended that all four test modules used in the field trial be retained for use in the main survey given that all four exhibited satisfactory validity, functioning, and measurement properties.

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 five CIL test modules (three trend modules, two newly developed modules) and two CT test modules, were then finalized.

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.

Field trial test design and content

Test design

In this report we refer to tasks, items, and score points when describing the ICILS tests of CIL and CT.

The term *task* refers to the instructions given to students and the actions required to complete them. As described previously, the ICILS test modules include a range of information-based response tasks, and skills, authoring, and coding tasks.

The term *item* refers to the variable or variables derived using the scored student responses to each task that were used to create the scales of CIL and CT and to measure student achievement against them. Achievement on tasks, such as information-based response tasks, and skills and coding tasks, was typically measured using one item per task. Achievement on the authoring tasks was measured using many items (scoring criteria) per task.

The term *score points* refers to the number of discrete non-zero score categories per item. The items associated with skills tasks typically elicited one score point each (e.g., correct = 1, incorrect = 0) whereas most of the items associated with information-based response tasks, authoring tasks,

and coding tasks elicited more than one score point (e.g., full credit = 3, partial credit (high) = 2, partial credit (low) = 1, and no credit = 0).

The CIL field trial test instrument consisted of five test modules with a total of 51 tasks which yielded data for 86 items (all authoring tasks and a small number of skills tasks were assessed using more than one criterion, with each criterion constituting an item). The selection of tasks, including their format, was determined by the nature of the content the tasks were assessing, the tasks' potential range of response types, and their role in the narrative flow of each module.

The ICILS research team had earlier decided not to have the same balance of task types and task formats within each module but rather to have a mix across the five modules of information-based response tasks, skills tasks, and authoring tasks. Table 2.3 shows the composition of the field trial CIL test modules by items derived from the different task types. Overall, 30 percent of the items were derived from information-based response tasks, 22 percent from skills task, and 45 percent from authoring tasks.

The CT field trial test instrument consisted of two test modules with a total of 21 tasks yielding data for 19 items (Table 2.4).[2]

Table 2.3: Field trial CIL test module composition by items derived from tasks

Module	Items by task format						
	Information-based response tasks			Skills tasks		Authoring tasks	Total
	Multiple choice	Constructed response	Drag and drop	Linear skills	Nonlinear skills	Large task criterion	
Band competition	1	2	2	2	3	7	**17**
Breathing	0	3	0	3	0	10	**16**
School trip	3	0	0	4	1	7	**15**
Board games	3	4	0	1	0	6	**14**
Recycling*	3	5	0	1	4	11	**24**
Total	**10**	**14**	**2**	**11**	**8**	**41**	**86**

Note: *The recycling module included a short note-taking task and a communication task. In this table both are classified as authoring (large) tasks.

Table 2.4: Field trial CT test module composition by items derived from tasks

Module	Items by task format						
	Information-based response tasks		Skills tasks				Total
	Multiple choice	Constructed response	Nonlinear systems transfer	Simulation	Algorithm construction	Algorithm debugging	
Automated bus	1	1	4	2	0	0	**8**
Farm drone	0	2	1	0	6	2	11
Total	**1**	**3**	**5**	**2**	**6**	**2**	**19**

2 Data for two tasks were not used in the field trial scaling analysis.

The test modules were delivered to students in a fully balanced complete rotation. Table 2.5 shows this design. In total, there were 60 different possible combinations of modules (20 combinations for students completing the test of CIL only and 40 combinations for students completing both the tests of CIL and CT).

There were 20 possible combinations of the two CIL modules selected from the pool of five to be administered to all students. Each module appeared in eight module combinations (four times in the first position and four times in the second position). In countries completing only the CIL component of ICILS, each student completed one of the 20 CIL module combinations followed by the student questionnaire.

In countries completing the CT option, each student completed one of the 20 possible CIL module combinations followed by the student questionnaire and then both CT modules presented in one of two possible sequences. This combination of 20 CIL module combinations each presented with one of two possible CT sequences resulted in 40 module combinations available for students to complete.

The term *test form* in Table 2.5 refers to each combination of modules that was used in the field trial.

Table 2.5: Field trial test form design and contents

CIL combination	CIL module position	
	1	2
1	B	H
2	B	S
3	B	G
4	B	R
5	H	S
6	H	G
7	H	R
8	S	G
9	S	R
10	G	R
11	H	B
12	S	B
13	G	B
14	R	B
15	S	H
16	G	H
17	R	H
18	G	S
19	R	S
20	R	G

Field trial test form design			
Test form	CIL combination	Questionnaire	CT order
1-20	1-20	Q	–
21-40	1-20	Q	AB:FD
41-60	1-20	Q	FD:AB

Module name key
B: Band competition
H: Breathing
S: School trip
G: Board games
R: Recycling
AB: Automated bus
FD: Farm drone

Field trial coverage of the CIL framework

All field trial items were developed according to and mapped against the ICILS CIL framework. Table 2.6 shows this mapping.

Table 2.6: Field trial CIL item mapping to the CIL framework

	CIL framework aspect	Items (n)	Items (%)	Score points (n)	Score points (%)
1.1	Foundations of computer use	4	5	5	4
1.2	Computer use conventions	9	10	10	8
2.1	Accessing and evaluating information	16	19	24	19
2.2	Managing information	9	10	11	9
3.1	Transforming information	14	16	23	18
3.2	Creating information	23	27	40	31
4.1	Sharing information	7	8	9	7
4.2	Using information responsibly and safely	4	5	6	5
		86	**100**	**128**	**100**

As stated in the ICILS 2018 assessment framework, "[t]he test design of ICILS was not planned to assess equal proportions of all aspects of the CIL construct, but rather to ensure some coverage of all aspects as part of an authentic set of assessment activities in context" (Fraillon et al. 2019, p. 54). The intention that the four strands would be adequately represented in the test was achieved. Twelve percent of score points related to Strand 1, 27 percent to Strand 2, 49 percent to Strand 3, and 12 percent to Strand 4. These proportions corresponded to the amount of time the ICILS students were expected to spend on each strand's complement of tasks. Aspects 2.1, 3.1, and 3.2 were assessed primarily via the large tasks at the end of each module, with students expected to spend roughly two thirds of their working time on these tasks.

Field trial coverage of the CT framework

All field trial items were developed according to, and mapped against, the ICILS CT framework. Table 2.7 shows this mapping.

Table 2.7: Field trial CT item mapping to the CT framework

	CT framework aspect	Items (n)	Items (%)	Score points (n)	Score points (%)
1.1	Knowing about and understanding digital systems	3	16	7	16
1.2	Formulating and analyzing problems	3	16	5	12
1.3	Collecting and representing relevant data	3	16	4	9
2.1	Planning and evaluating solutions	4	21	10	23
2.2	Developing algorithms, programs, and interfaces	6	32	17	40
		19	**100**	**43**	**100**

As stated in the ICILS assessment framework, "[t]he test design of ICILS 2018 was not planned to assess equal proportions of all aspects of the CT construct, but rather to ensure some coverage of all aspects as part of an authentic set of assessment activities in context" (Fraillon et al. 2019, p. 54). Table 2.6 shows that, while there was a similar number of items addressing each of Strands 1 and 2 of the CT framework, a majority (63%) of score points collected related to Strand 2. This is because the quality of students' responses to algorithm construction and algorithm debugging tasks (in the farm drone module) was typically assessed using partial credit criteria (of up to three score points each) that combined measures of the accuracy and elegance of students' coding solutions for each task. Detailed information about scoring the algorithm construction and algorithm debugging tasks has been included in the section in this chapter describing the main survey test design and content.

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 staff 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.

Also as mentioned previously, ISC staff 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. After consulting with NRCs, the ISC decided to retain all five CIL field-trial test modules and the two CT test modules for use in the main survey.

Feedback from NRCs and observers of the field trial suggested that the total testing time was long for students completing both the CIL and CT tests and that many students were finishing the automated bus module in far less than the allocated 30 minutes. As a result of this, it was agreed that the time allocated to each CT module would be reduced to 25 minutes and that the farm drone module would be shortened by removing two final constructed response items that required students to provide reflections on coding solutions. It was also agreed that the farm drone module would benefit from an increase in the number of algorithm debugging tasks. Two algorithm construction tasks from the field trial were converted to algorithm debugging tasks for use in the main survey.

Main survey test design and content

Test design

The main survey test instrument consisted of five test modules with a total of 46 tasks which yielded data for 81 items. Some of these tasks generated a number of score points based on the criteria that were applied to the large tasks. Table 2.8 shows the composition of the main survey test modules by items derived from the different task types. The items shown in Table 2.8 are those that were used in the scaling and analysis of the ICILS 2018 main survey CIL test data. Overall, 27 percent of the items were derived from information-based response tasks, 22 percent from skills task, and 51 percent from authoring tasks.

The CT main survey test instrument consisted of two test modules with a total of 17 tasks which yielded data for 17 items. The CT test emphasized the application of skills with 15 of the 17 items derived from skills tasks (Table 2.9).

Students received the test modules in the same fully balanced complete rotation that was used in the field trial. Table 2.10 shows this design for the main survey. As before, the term *test form* refers to each combination of modules used in the main survey.

Table 2.8: Main survey CIL test module composition by items derived from tasks

Module	Items by task format						
	Information-based response tasks			Skills tasks		Authoring tasks	Total
	Multiple choice	Constructed response	Drag and drop	Linear skills	Nonlinear skills	Large task criterion	
Band competition	1	2	2	2	3	7	**17**
Breathing	0	2	0	3	0	10	**15**
School trip	1	0	0	4	1	7	**13**
Board games	2	4	0	1	1	5	**13**
Recycling*	3	5	0	1	2	12	**23**
Total	**7**	**13**	**2**	**11**	**7**	**41**	**81**

Note: *The recycling module included a short note-taking task and a communication task. In this table both are classified as large tasks.

Table 2.9: Main survey CT test module composition by items derived from tasks

Module	Items by task format						
	Information-based response tasks		Skills tasks				Total
	Multiple choice	Constructed response	Nonlinear systems transfer	Simulation	Algorithm construction	Algorithm debugging	
Automated bus	1	1	4	2	0	0	**8**
Farm drone	0	0	1	0	5	3	**9**
Total	**1**	**1**	**5**	**2**	**5**	**3**	**17**

Table 2.10: Main survey test form design and contents

CIL combination	CIL module position	
	1	2
1	B	H
2	B	S
3	B	G
4	B	R
5	H	S
6	H	G
7	H	R
8	S	G
9	S	R
10	G	R
11	H	B
12	S	B
13	G	B
14	R	B
15	S	H
16	G	H
17	R	H
18	G	S
19	R	S
20	R	G

Main survey test form design			
Test form	CIL combination	Questionnaire	CT order
1-20	1-20	Q	-
21-40	1-20	Q	AB:FD
41-60	1-20	Q	FD:AB

Module name key
B: Band competition
H: Breathing
S: School trip
G: Board games
R: Recycling
AB: Automated bus
FD: Farm drone

Main survey coverage of the CIL framework

All main survey items were developed according to and mapped against the ICILS CIL framework. Table 2.11 shows this mapping.

Table 2.11: Main survey CIL item mapping to the CIL framework

	CIL framework aspect	Items (n)	Items (%)	Score points (n)	Score points (%)
1.1	Foundations of computer use	2	2	2	2
1.2	Computer use conventions	11	14	13	13
2.1	Accessing and evaluating information	14	17	16	16
2.2	Managing information	8	10	8	8
3.1	Transforming information	15	19	20	20
3.2	Creating information	21	26	31	30
4.1	Sharing information	7	9	8	8
4.2	Using information responsibly and safely	3	4	4	4
		81	**100**	**102**	**100**

A comparison of Tables 2.6 and 2.11 reveals that the final test instrument provided very similar CIL framework coverage to that of the field trial instrument.

The 81 main survey items yielded 102 score points for inclusion in the item analysis and scaling. Overall 15 percent of score points were derived from items corresponding to Strand 1, 24 percent to Strand 2, 50 percent to Strand 3, and 12 percent to Strand 4.

Main survey coverage of the CT framework

All main survey items were developed according to and mapped against the ICILS CT framework. Table 2.12 shows this mapping.

Table 2.12: Main survey CT item mapping to the CT framework

	CIL framework aspect	Items (n)	Items (%)	Score points (n)	Score points (%)
1.1	Knowing about and understanding digital systems	3	18	7	18
1.2	Formulating and analyzing problems	2	12	4	10
1.3	Collecting and representing relevant data	3	18	5	13
2.1	Planning and evaluating solutions	5	29	12	31
2.2	Developing algorithms, programs, and interfaces	4	24	11	28
		17	**100**	**39**	**100**

A comparison of Tables 2.7 and 2.12 reveals that the final test instrument provided similar CT framework coverage to that of the field trial instrument. The relative decrease in Strand 2.2 coverage and increase in Strand 2.1 coverage between the field trial and main survey instrument was largely a result of a conversion of two algorithm construction tasks in the field trial to become algorithm debugging tasks in the main survey. Other small changes in the coverage by aspect related to the removal of two constructed response evaluation tasks from the farm drone task following the field trial.

The 17 main survey items yielded 39 score points for inclusion in the item analysis and scaling. Overall 41 percent and 59 percent of score points were derived from items corresponding to Strands 1 and 2 respectively.

Scoring the main survey CT algorithm construction and algorithm debugging tasks

CT items

The CT test was new to ICILS and represented the first time CT has been assessed in a cross-national large-scale assessment content. The quality of students' responses to visual coding tasks has not previously been assessed in this context and consequently we have included this section in the technical report to explain the conceptual basis and operationalization of the scoring of students' visual code algorithms in ICILS 2018.

The algorithm construction tasks required students to select commands (available as visual code blocks) and place them in sequence in order for the farm drone to complete specified tasks. The tasks included any or all of the following actions: having the drone move to specified locations on a grid, having the drone drop any of water, seed, or fertilizer, and applying conditional logic relating to the size of the crops on which an action may be conducted (using a simple but configurable "if…

do" command). After two simple warm-up tasks, students were also introduced to and able to make use of a configurable "repeat" command used to complete multiple iterations of commands.

The algorithm debugging tasks required students edit an existing configuration of commands in order for the farm drone to complete specified tasks. The algorithm debugging tasks were developed such that one or two minor modifications to the command configuration could result in the drone completing the actions as required. However, there were no restrictions in the debugging tasks on the nature of the changes students could make to the commands and their configuration. Students could, if they wished, remove all the existing commands and develop a completely new set of code.

The tasks became progressively more complex as students worked through the module. Task complexity related to the number of actions needed to be completed by the drone and the number, type, and configuration of targets onto which the drone needed to drop any of water, seed, or fertilizer.

The quality of students' coding solutions to both algorithm construction and debugging tasks was conceptualized in terms of two main criteria. The first criterion related to the *correctness* of any given solution and the second was the *efficiency* (or *elegance*) of the solution. Ultimately we derived measures for each criterion and then combined these into a single measure of the quality of the solution to each coding task. As we defined, specified, and operationalized the measures for each criterion, it was possible to have the computer-based delivery system generate the relevant data for each student response. So while the identification of variables and scores to measure the quality of coding were determined by the research team, the individual scores for student responses were ultimately generated through the test delivery system. The method for scoring each criterion and then combining the scores to form a single score for each task is described below.

Scoring the correctness of coding solutions

The correctness of a solution was characterized by the degree to which the behavior of the drone in response to a student's coding solution matched the required behavior of the drone specified in the task instructions. In a small number of simple tasks, the instruction was that the drone fly to a specified location. For these tasks, the correctness was measured only in terms of whether or not the drone ended up at the specified location. For the majority of tasks the drone was required to both fly to specified locations and drop any of water, seed, or fertilizer on these targets. For these tasks, the correctness of the solution was initially scored according to three sub-criteria that comprised two variables (*correct targets* and *incorrect targets*):

i. the number of targets with the correct materials on them (correct targets: 2 score points per target with correct materials)
ii. the number of targets with incorrect materials on them (correct targets: 1 score point per target with incorrect materials)
iii. whether any materials were dropped on a square that was not a target (incorrect targets: 1 score point for no materials on any incorrect targets, 0 score points for any materials on any incorrect target).

Each response was first scored for each variable (correct targets and incorrect targets). Following this, the frequencies of the scores were examined together with consideration of the descriptions of combinations of these scores. This was an iterative process used to establish a scoring logic for the correctness on each task. This scoring logic included combining the correct targets and incorrect targets scores into a single initial combined correctness score for each task. Once the scoring logic for the correctness of each task was established, we completed a Rasch item response theory scaling analysis (Rasch 1960) (see Chapter 11 for further detail) to establish the measurement viability of the scoring logic and, where appropriate, we recoded scoring categories for selected

items. Table 2.13 shows an example operationalization of this scoring logic for a task, together with the recoding of categories following initial scaling. For each task, the score for the completely correct response remained as a unique and highest score category. The recoding combination of score categories for partially correct responses varied across tasks and was informed by the response frequencies, the substantive differences in response quality, and the degree to which these differences were reflected in the student data.

Table 2.13: 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 the efficiency of coding solutions

The second criterion used to score the student coding responses related to the efficiency (or elegance) of students' responses. We explored two main approaches when considering the efficiency of the code in students' responses. The first was a simple count of the number of commands in each response. For each task the most efficient response was defined as one that included the smallest number of commands needed to complete a completely correct response (for each task in the farm drone module students were instructed to use as few commands as possible in their solution). 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.

The second approach was to examine the degree to which students used embedded logic within their code solutions, such as loops within loops. This was considered for its potential to represent the elegance of the substance of a coding solution. For the more complex tasks it was possible to observe some variation in the degree to which students embedded commands and in an early phase of the analysis we included an elegance variable derived from the degree to which embedded commands were used.

Once we had established efficiency scores for responses to each task and elegance scores for responses to the more complex tasks, we included these data in the scaling model (see Chapter

11). We found that the data from the elegance variable did not contribute to the quality of the measurement of student CT and at this point decided to use only the efficiency data to contribute to the scoring. Table 2.14 shows an example operationalization of the scoring logic in establishing the efficiency scores for a task, together with the recoding of categories following initial scaling.

Table 2.14: 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 the correctness and efficiency scores to establish a single score for each coding task

The final step of establishing scores for each coding task was to combine the correctness and efficiency scores. This was done by using the efficiency score for each task 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.

The correctness scores for all partially correct or incorrect responses were unchanged when included in the correctness and efficiency composite variable. Table 2.15 shows an example of how correctness and efficiency scores were combined to form the single composite variable.

Released CIL test module and CT tasks

One CIL test module, *band competition*, has been released since publication of the ICILS international report (Fraillon et al. 2020). This module required students to work on a sequence of tasks associated with planning a website to promote a band competition within a school. The large task for this module required students to add content to a page on the band competition website.

The large task in this module presented students with a description of the task details as well as information about how the task would be assessed. The description was followed by a short video designed to familiarize students with the task and highlight the main features of the software they would need to use to complete the task. A detailed description of the module appears on pages 60 to 74 of the ICILS 2018 international report (Fraillon et al. 2020).

Table 2.15: 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 1
0	N/A	Fewer than 2 targets met	0

Four tasks from the CT test have also been released. Two tasks were taken from the automated bus module and show the use of a simulation configuration of a decision tree. Two tasks were taken from the farm drone module and show algorithm construction and algorithm debugging. Detailed descriptions of these tasks appear on pages 94 to 101 of the ICILS international report (Fraillon et al. 2020).

References

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030193881

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*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030387808

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

CHAPTER 3:

Computer-based assessment systems

Julian Fraillon, Ralph Carstens, and Sebastian Meyer

Introduction

ICILS 2018 collects student achievement and questionnaire data on computer rather than on paper. This chapter describes the key aspects of the computer-based test delivery system used in ICILS 2018. It also details some of the challenges of using computer-based systems to collect student data in a crossnational large-scale assessment such as ICILS 2018.

The focus of the chapter is on the overall approach, architecture, and design of the computer-based assessment (CBA) system suite as well as the relationship of these to other technical systems used for the survey operations. Details and procedural aspects are provided in other 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.

ICILS 2018 computer-based components and architecture

Pre-existing components

IEA has been using computer-based systems for a number of years in order to organize and support countries to coordinate field operations and collect questionnaire data from teachers and schools. These systems, developed by IEA, include the following:

- The *IEA Windows Within-School Sampling Software* (IEA WinW3S), which supports countries to manage within-school sampling and test administration procedures;
- The *IEA Online Survey System* (IEA OSS), which supports the translation, adaptation, and subsequent delivery of computer-based questionnaire material;
- The *IEA Data Management Expert* (IEA DME), which is used to capture data from paper-based questionnaire material and to integrate and verify national databases;
- The *IEA Coding Expert*, which is used to code the student responses with respect to parental occupation; and
- The *IEA Translation System*, which is used to translate and review the school, ICT coordinator, and teacher questionnaires online.

All five software components were used in ICILS 2018 once they had undergone significant customizations and/or extension so that they would suit the specific study context.

Data from paper- and computer-based test and questionnaire components were transformed and integrated in the pre-existing IEA Data Processing Expert software. Chapter 10 provides more information on this process.

As in ICILS 2013, RM Results (previously SoNET Systems) provided the software components directly relating to the collection of student data via computer for ICILS 2018. These software components are part of a broader suite known as AssessmentMaster (AM). The software modules specifically relating to the delivery of the ICILS 2018 student test and questionnaire include an administration module (AM Manager), a translation module (AM Designer), a delivery engine module (AM Examiner), and a scoring module (AM Marker).

Three ICILS 2018 student test modules were already administered in ICILS 2013. They are usually referred to as trend modules.

Components developed for ICILS 2018

Apart from general updates and improvements in all software packages used in ICILS 2018, four new test modules were introduced. Two modules for the computer and information literacy (CIL) test and the two new modules comprising the computational thinking (CT) test.

Procedures and software user manuals

A series of detailed manuals guides the work of national center staff during IEA studies. These manuals are known as Survey Operations Procedures (SOP) manuals. IEA, in close cooperation with the entire international team, has developed and refined instructions and guidelines for IEA WinW3S, IEA OSS, IEA DME, and the IEA Translation System and included these in the suite of SOP materials. Each user manual is tailored to the needs of each project (see Chapter 8 for details).

During ICILS 2018, a separate user manual was developed for the translation module, while instructions on how to use the scoring and administration modules were included within the SOP materials. The manuals relating to the delivery engine module were incorporated in the manuals for school coordinators and test administrators who were administering the tests in schools.

ICILS 2018 system architecture

Table 3.1 provides an overview of the ICILS 2018 computer-based system and its supporting and accompanying systems. The table also shows how these systems interfaced and interacted. The table is organized by the major components of the system and the ICILS 2018 target populations. As can be seen, ICILS 2018 used a mixture of pre-existing and newly developed and/or customized components and tools to support the preparation, administration, and post-processing of the study.

Developing the computer-based delivery platform for ICILS 2018

Developing the delivery engine

Development of the computer-based delivery platform took place in parallel with test development (see Chapter 2 for details of the test development). The three trend test modules, i.e., the ones administered in ICILS 2013, were already available and largely ready-to-use. In response to some changes in the delivery system, AM Examiner, RM Results updated aspects of the backend functioning of the trend modules. These changes did not affect the user-experience of the modules in 2018 in comparsion to 2013. Two new CIL modules and two CT modules were developed for ICILS 2018. Once the new test modules' storyboards had been completed, they were sent to RM Results for authoring into the delivery platform. This process was an iterative one wherein the necessary functionality of each task in the test was reviewed and refined once it had been enacted. Each task underwent many such iterations.

The test delivery engine was web-based, which meant that technically the tests could be delivered over the internet. However, the ICILS 2018 research team decided, for operational reasons, that the tests should be administered on USB drives (one per computer), each containing a self-contained web-server to deliver the web-based test content. Although the USB-based delivery engine was Windows-based, it could be run using some forms of Windows emulation on Mac OS X or Linux.

The USB hosted a portable version of the Mozilla Firefox browser. Accordingly, all ICILS 2018 materials were developed, rendered, and tested using Mozilla Firefox and Google Chrome. The research team recommended that the translation, scoring, and administration modules be accessed only through Firefox (although Google Chrome was also an option). As part of the field-operation procedures, national centers needed to ensure that session data from each USB drive were uploaded to a central database as soon as practicable after each test session.

Table 3.1: ICILS computer-based or computer-supported systems and operations

Process	Student test and questionnaire	Principal, ICT coordinator, and teacher questionnaires (online)	Principal, ICT coordinator, and teacher questionnaires (paper)
Authoring/master instruments	AssessmentMaster authoring from storyboards	IEA OSS admin module, transfer from IEA Translation System	Microsoft Word, direct authoring of paper-based questionnaires
Translation	AM Designer, web-based	IEA Translation System	Microsoft Word, desktop-based
Sample selection and ID provisioning	IEA WinW3S		
Delivery systems	AM Examiner	IEA OSS delivery module, web-based	Personalized questionnaire prints
Initialization	Student ID, password, and language information from IEA WinW3S	Respondent ID and password from IEA WinW3S	Labels with respondent ID from IEA WinW3S
Primary delivery mode	USB sticks: Windows-based on local school computers; proctored	Any internet browser: self-administered	Paper: self-administered
Alternative delivery mode(s)	Laptop server mode or carry-in laptop sets (where school infrastructure insufficient)	Paper questionnaires (where infrastructure insufficient)	N/A
Data capture	Directly onto USB sticks; alternatively, local laptop servers (student by student)	Directly into central database (all respondents)	Manual (human) data using IEA DME software after administration (all respondents)
Data merging (across respondents)	Upload to AM Manager	N/A	Merging from multiple IEA DME databases (where used)
Data scoring	AM Marker	N/A	N/A
Data management	IEA WinW3S; crosscheck of expected versus available data		

An alternative delivery option was made available for use in the main survey. Under this system, the delivery engine was installed on a notebook computer connected to a school LAN. The school could then access the test locally over the school LAN via Mozilla Firefox or Google Chrome.

School coordinators were provided with the following list of minimum specifications for computers to run the ICILS student test from USB sticks.

- Screen resolution of at least 1024x768;
- One of the following operating systems: Windows XP/Service Pack 3/Windows Vista/ Windows 7/Windows 8/Windows 10;
- Display (fonts) set to the optimal size (minimum: 100%; the system font size can be set from the computer's Control Panel/Appearance and Personalization/Display); and
- A USB Port 2.0 or higher.

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

Requirements for the **server computer:**

- Screen resolution of at least 1024x768;
- One of the following operating systems: Windows 7/Windows 8/Windows 10* (*Windows 10 was recommended for the best experience);
- A recommended CPU of 1500 MHZ;
- An installed System Memory of at least 4 GB;
- At least 10 GB of free storage space on the hard drive;
- LAN or Wi-Fi connectivity; and
- Internet connectivity (only required for data upload).

Requirements for the **client computers**:

- Screen resolution of at least 1024x768;
- Connectivity to the same LAN or Wi-Fi network as the server computer;
- Recent version (less than 12 months old) of the Mozilla Firefox or Google Chrome browser installed; and
- Display (fonts) set to the optimal size (minimum: 100%; the system font size can be set from the computer's Control Panel/Appearance and Personalization/Display).

The delivery module was developed to include 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 large/authoring tasks (live software applications with functionality to add, edit, and format text in a range of formats);
- Display of closed web environments with the facility to display multiple pages and navigate between and within pages;
- Capture of all student final responses to each task;
- Capture of time taken on each task;
- Facility to score student responses to skills tasks;
- A progress display (number of tasks completed and to come per module) for students; and
- A countdown timer for students per module.

Developing the translation systems

In ICILS 2018 two different translation systems were used: the IEA Translation System and AM Designer. This corresponds to the fact that two delivery systems for the electronic administrations of ICILS 2018 instruments were also used. The school, teacher, and ICT coordinator questionnaires were administered via the IEA OSS (if not administered on paper) while the student tests and questionnaires were administered using AM Examiner.

Both translation systems are web-based applications accessed through a web-browser. Users needed the following in order to access it:

- A computer with a broadband or equivalent high-speed internet connection; and
- A web browser, e.g., Mozilla Firefox or Google Chrome.

Development of the translation systems provided 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 and across tasks;
- Ability to bulk-import the field trial translations so that they could be used as a starting point for main survey translations; and
- Opportunity for translators, reviewers, and verifiers to monitor the progress of task completion (i.e., translation state).

Both systems included the functionalities listed above. The school, teacher, and ICT coordinator questionnaires were translated and verified within the IEA Translation System while the student instruments (test and questionnaires) were translated and verified within AM Designer.

Developing the scoring module

The ICILS 2018 scoring module, AM Marker, was an adaptation of the ICILS 2013 scoring module. This adaptation included the development of some new features for use in ICILS 2018. It also included replacing (where possible) text in the user interface with icons so that the user interface did not need to be translated. (ICILS 2018 did not require scorers to be proficient in English.)

The web-based scoring application could be accessed through a web-browser. In order to access the application, users needed:

- A computer with a broadband or equivalent high-speed internet connection; and
- A recent version (12 months old or less) of Mozilla Firefox or Google Chrome.

AM Marker was developed to include the following features:

- Selective functionality relating to the role of the user (e.g., administrator, scoring trainer, team leader, scorer);
- Facility to specify a proportion (in ICILS 2018, 20%) of student responses to be blind double-scored for the purpose of monitoring inter-rater reliability; and
- Capacity to begin scoring before all student responses had been uploaded to the system (i.e., before completion of data collection in schools) without compromising the double-scoring procedure.

Scorers were able to:

- View tasks (as they appeared to students in the test) along with student responses on screen;
- Enter a score for each student for each task;
- Flag pieces of work for follow-up with a more senior staff member;
- Navigate back to previously scored pieces of work and amend scores;

- View large tasks with full functionality; and
- Enter a "training" mode to score pre-scored work and receive feedback on scoring accuracy.

In addition, team leaders could:

- Review (check-score and amend) scorers' scores; and
- Monitor and respond to flagged pieces of work.

In addition, scoring trainers could:

- Select, order, and annotate responses for use in scorer training.

In addition, scoring administrators could:

- Allocate scorers to teams; and
- View reports of interscorer reliability.

Developing the administration module

The ICILS 2018 administration module, AM Manager, was developed to (i) support national center staff monitor the progress of test sessions (i.e., data upload), create user accounts, and define roles for users of the other modules (scoring and translation), and (ii) export test-session data for importing into IEA WinW3S.

The administration module was a web-based application that users accessed through a web-browser. In order to access the application, users needed:

- A computer with a broadband or equivalent high-speed internet connection; and
- A recent version (12 months old or less) of Mozilla Firefox or Google Chrome.

Development of the administration module focused on ensuring that users could:

- Create user accounts and allocate roles;
- Download the software image of the ICILS 2018 test (including the test delivery engine);
- View national test session details;
- Monitor national test session status; and
- Export test session data for IEA WinW3S.

Challenges with computer-based delivery in ICILS 2018

Successful collection of data in large-scale computer-based surveys relies on individual participants being able to provide responses on computers in controlled uniform 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.

Uniformity of test presentation and administration is easily assured in a printed test, but CBAs present additional challenges arising out of variations in presentation, administration, and the utilized infrastructure. These issues can influence respondents' performance and are 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 ICILS 2018 test-delivery platform, 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 2018 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 2018 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 computer-based system components (as shown in Table 3.1). As is the case with any system comprising interconnected components with a range of specific functions, the interfaces between each component are potential points of failure. To ensure that the system worked successfully, survey 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 national survey coordinators and their staff classified as technical problems during administration. Nonresponse 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 surveys. Those responsible for implementing a survey 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 2018, the inspection of unweighted yet cleaned data from all 14 participating education systems (52,625 sampled students) 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 dispositions for all sampled students (so excluding any noncooperating 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	1.3%	
Parental permission denied	2.6%	4.2%
Absent	6.4%	7.0%
Incompatible or failed equipment before assessment	0.1%	
Technical failure during assessment	0.3%	0.1%
USB stick lost or upload failed after assessment	0.1%	
Participated	89.2%	87.2%
Total	100%	

Table 3.2 shows that test and questionnaire data were collected from 89.2 percent and 87.2 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 (0.6% more students were absent for the questionnaire than the test).
2. In some countries parents gave permission for their children to complete the test but not the questionnaire (resulting in a further absentee difference of 1.6%).

In some cases the test adminstrators managed to resolve technical issues for the questionnaire session where time is not as critical as during the test session (0.2% fewer technical failures during the questionnaire session).

Roughly 7.7 percent of sampled students overall had either permanently left school (1.3%) or were absent on the day of testing (6.4%).

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. 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.

An initial review of a survey activities questionnaire completed by national research coordinators (NRCs) indicated no systematic failures or problems with the testing system, although technical issues were observed in comparable proportions before, during, and after the assessment. However, none of the NRCs indicated or reported a lack of suitable delivery options (USB plus the local server alternative) as responsible for instances of school nonresponse. The fact that each assessment was executed on a different computer (except for carry-in laptops that were uniformly configured) may alone account for the small number of apparently unconnected errors observed.

Some NRCs did report the general challenge of attaining school cooperation and participation in light of survey fatigue, industrial actions, or other factors unrelated to the assessment and its data collection mode. Among those schools that agreed to participate, the proportion of students not participating because of (i) denied parental permissions, (ii) school leaving, or (iii) other types of incidental absences (e.g., sick-listing) accounted for a level of nonresponse many times larger than that due to technical issues.

Compared to 2013 the technical problems, cases involving lost USB sticks, corrupted data, or data that could not be uploaded seems to have been only a minor issue in ICILS 2018 (0.1% of all sampled students) while in ICILS 2013 this constituted the largest factor (0.9%).

CHAPTER 4:

ICILS 2018 questionnaire development

John Ainley and Wolfram Schulz

Introduction

In this chapter, we describe the development of the ICILS 2018 questionnaires for students, teachers, school principals, ICT coordinators, and national research centers.

The student questionnaire was designed to gather information about students' personal and home background as well as use of and familiarity with computers and computing. It included questions relating to students' background characteristics, their experience and use of computers and ICT to complete a range of different tasks in school and out of school, and their attitudes toward the use of computers and ICT in society. The teacher questionnaire was designed to gather teachers' perspectives on the school environment for ICT use, their confidence in using ICT, their attitudes regarding the use of ICT in teaching, and their use of ICT for teaching and learning.

There were two school questionnaires: one completed by the school principal and one completed by the ICT coordinator. School principals were asked to report on ICT use in learning and teaching at their school, school characteristics, and teachers' professional development in using ICT at school. The ICT coordinator questionnaire included questions about the availability of ICT resources at school (e.g., infrastructure, hardware, and software) as well as pedagogical support for the use of ICT.

An online questionnaire, the national contexts survey, for the national research coordinators (NRCs) was designed to collect contextual information at the national (or sub-national) level about the characteristics of education systems, plans, policies for using ICT in education, ICT and student learning at lower-secondary level, ICT as part of teachers' professional development, and the existence of ICT-based learning and administration systems in the country.

The conceptual framework used to guide questionnaire development

The assessment framework provided a conceptual underpinning for the development of the international instrumentation for ICILS (Fraillon et al. 2019). The assessment framework consisted of three parts:

- *The computer and information literacy framework*: This outlined the measure of computer and information literacy (CIL) that was addressed through a cognitive test and those parts of the student questionnaire designed to measure student perceptions regarding CIL.
- *The computational thinking framework*: This outlined the measure of computational thinking (CT) that was addressed through a cognitive test and parts of the student questionnaire designed to measure student perceptions related to CT.
- *The contextual framework*: This mapped the context factors expected to influence outcomes, explain variation in those outcomes, and map student usage of ICT in school and out-of-school settings.

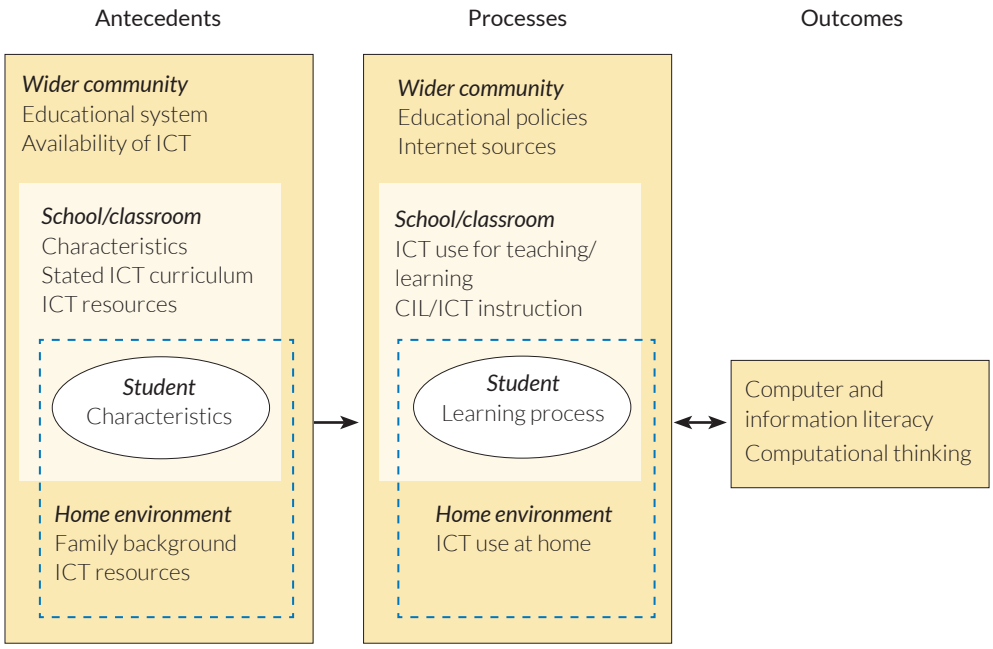
The contextual framework identified the variables that reflect the environment in which learning CIL and CT takes place. It assumes that young people develop their understandings of ICT through activities and experiences that take place in schools and classrooms, and in their homes and other places outside school.

Students' development of CIL and CT is influenced by what happens in their schools and classrooms (the instruction they receive, the availability and use of ICT resources in schools, ICT use for teaching and learning), their home environments (socioeconomic background, availability and

use of ICT at home), and their individual characteristics. Factors related to each of these different levels shape the way students respond to learning about computers and computing.

Contextual influences on CIL and CT learning are conceived as either antecedents or processes (Figure 4.1). Antecedents refer to the general background that affects how CIL and CT learning takes place (e.g., through context factors such as ICT provision and curricular policies that shape how learning about ICT is provided). Process-related variables are those factors shaping CIL and CT learning more directly (e.g., the extent of opportunities for CIL and CT learning during class, teacher attitudes toward ICT for study tasks, and students' computer use at home).

Figure 4.1: Contexts for ICILS 2018 CIL/CT learning outcomes



In reference to this conceptual framework structure, variables collected through contextual instruments with examples of different types of measures are displayed in Table 4.1 below, where columns contain antecedents and processes and rows the four levels. The student questionnaire collected data on student experience, use, and perceptions of ICT as well as contextual factors at the individual (either school or home) level. The teacher, principal, and ICT coordinator questionnaires focused on gathering data to be used at the school level while the national contexts survey and published sources provided variables at the system or national level.

Table 4.1: Mapping of variables to contextual framework with examples

Level of ...	Antecedents	Processes
Wider community	**NCS & other sources:** Structure of education Accessibility of ICT	**NCS & other sources:** Role of ICT in curriculum
School/classroom	**PrQ, ICQ, & TQ:** School characteristics ICT resources	**PrQ, ICQ, TQ, & StQ:** ICT use in teaching and learning CIL/CT instruction
Student	**StQ:** Gender Age	**StQ:** ICT activities Use of ICT CIL/CT
Home environment	**StQ:** Parent socioeconomic status Home ICT resources	**StQ:** Learning about ICT at home

Notes: NCS = national contexts survey; PrQ = principal questionnaire; ICQ = ICT coordinator questionnaire; TQ = teacher questionnaire; StQ = student questionnaire.

Development of the ICILS context questionnaires

The international study center (ISC) at ACER coordinated the development and implementation of the ICILS 2018 context questionnaires for students, teachers, and schools. Several national centers also provided suggestions for additional questionnaire item material. The development work included extensive reviews and discussions at different stages of the process with experts and national centers.

The development process for the student, teacher, and school questionnaires followed a sequence which paralleled that for the CIL and CT tests. Specifically, the questionnaire development involved three phases:

- *Phase 1:* From February 2015 to September 2016 field trial material was developed with the content being guided by the assessment framework. It also included various rounds of consultations and reviews by NRCs (at meetings in March 2015, February 2016, and September 2016 as well as via online reviews in June and July 2016).
- *Phase 2:* Preparations for the field trial took place from November 2016 to May 2017. From May to June 2017 the international field trial was conducted in 14 participating ICILS countries. Subsequent analyses of field trial data took place from July to August 2017 to inform judgements about the suitability of questionnaire material for the main survey.
- *Phase 3:* The final phase focused on the selection of items for the main survey. From September to October 2017, ISC staff discussed the field trial results with staff in the national centers (including a meeting with NRCs in September 2017). Phase 3 concluded with a final selection of main survey items. From November 2017 to February 2018 survey materials were translated and the translations were verified.

The procedures and criteria used to review the field trial material and results were the same for the student, teacher, and school questionnaires. During instrument development, particular attention was paid to the appropriateness of questionnaire material for the large variety of national contexts 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 regard to the ICILS 2018 assessment framework;
- Appropriateness for the national contexts of the participating countries; and
- Psychometric properties of items designed to measure latent traits.

Because the national contexts survey did not include a field trial (see section below), the procedures used to develop it differed from those used for other context questionnaires.

Development of the student questionnaire

Students completed the ICILS student questionnaire on computer after they had completed the CIL test. In countries taking part in the CT option, the student questionnaire was completed after the CIL test and before the CT test. The student questionnaire collected information about the students' characteristics and their home background as well as the students' use of and familiarity with ICT. ICILS researchers at ACER coordinated the development and implementation of the student questionnaire.

The ICILS field trial student questionnaire material included a total of 33 questions (30 were international core questions and three were optional with 171 items (148 were core and 23 were optional). To reduce the length of the assessment time per student while trialing a larger pool of item material, the field trial student questionnaire was administered in three different forms. We allocated these forms in a way so that it was possible to analyze all possible combinations of item sets and scales. The field trial questionnaire was completed by samples consisting of 6695 students from 350 schools. On average in each country, 478 students provided data for the field trial analysis.

The analyses of the field trial data provided empirical evidence on the quality of the item material and informed the selection of the main survey material. ISC staff particularly emphasized the need to investigate the cross-national validity of the measures derived from the ICILS questionnaires (see examples from the International Civic and Citizenship Education Study [ICCS] 2009 in Schulz 2009). ISC staff analyzed field-trial data in terms of the distributions of responses (to check for skewness, etc.), the dimensionality and structure of the scales representing constructs and based on items (using exploratory and confirmatory factor analyses), and the reliabilities of the scales based on those sets of items designed to measure the corresponding constructs. These analyses were conducted for each participating education system in order to check that the measurement of instruments was equivalent across national contexts. Staff then discussed the field trial outcomes and the proposed draft student questionnaire for the main survey with NRCs. Their feedback helped the ISC staff to select item material for main survey instruments used in the final data-collection stage of ICILS 2018.

The final international student questionnaire consisted of 31 questions (28 were core questions and three were optional for countries) containing 135 items (123 were core and 12 were optional for countries), to be completed within the targeted time of 20 to 25 minutes. Twenty-two of the items (including one optional item) were designed to capture student-background information (including ICT experience) and 113 were designed to measure students' use of and familiarity with ICT (including 11 optional items). 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 education.
- *Your home and your family:* These questions focused on characteristics of the students' homes (including ICT resources) and their parents' occupations and educational backgrounds.

- *Your use of ICT:* These questions asked students to report on their experience with ICT, who taught them various ICT skills, their use at different locations, and their frequency of use of ICT for different activities.
- *Using ICT for school:* These questions asked students to indicate their frequency of use of ICT for school-related purposes in general and in specified learning areas, their use of specified ICT tools, and the extent to which they had learned at school how to do specified ICT tasks.
- *Your thoughts about using and learning about ICT:* These questions were designed to measure students' beliefs about their ability to do ICT tasks (ICT self-efficacy), their attitudes toward ICT in society, and the extent to which they had learned about CT-related tasks.

The student questionnaire items were designed to be administered on 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).

Development of the teacher questionnaire

The teacher questionnaire was designed to collect contextual information about school and classroom contexts for ICT learning, use of ICT for teaching and learning, teacher views on the pedagogical use of ICT, and their confidence in using computers. ISC staff at ACER coordinated development and implementation of the questionnaire and asked external experts and experts from national centers to review it at different stages of the study. The questionnaire was administered primarily through an online system but with provision for paper-based delivery in case teachers were unable or unwilling to complete the questionnaire online.

Under the assumption that teaching staff constitute an important factor in determining the extent to which ICT is used within the school context, the ISC staff responsible for designing the teacher questionnaire directed the questionnaire at all teachers teaching at the target grade (typically grade 8). They also designed the questionnaire so that teachers could complete it in about 30 minutes.

The questionnaire included a question about whether teachers used ICT in their teaching and learning. Those teachers who said yes were asked to provide information relating to a "reference class" about the extent of ICT use in this class for different purposes and activities. The reference class was specified for teachers as the first regular class at the target grade they had taught on or after the last Tuesday before answering the questionnaire.

The field trial teacher questionnaire consisted of 20 questions (including two optional questions) with a total of 159 items (including 22 optional items). It was administered to teachers from all subjects teaching at the target grade in the schools selected for the field trial. The field trial teacher sample consisted of 4236 teachers from 365 schools in 14 participating countries. On average, the field trial teacher samples consisted of about 300 teachers in each participating country. The analyses from the field trial teacher questionnaire data provided a basis for the selection of item material for the main survey.

The final main survey teacher questionnaire consisted of 18 questions with 116 items (including 10 items that were optional to countries) and was divided into the following five sections:

- *About you:* These questions concerned teachers' background characteristics and the subjects that they taught.

- *Your use of ICT:* These questions focused on teachers' experience of ICT, their frequency of use of ICT, and their confidence in performing ICT tasks.
- *Your use of ICT in teaching:* These questions asked teachers to name a reference class, provide information about the subject taught in that class, and state whether they used ICT for teaching and learning activities in this class. Those teachers who said they used ICT were asked to indicate the emphasis given to the development of CIL- and CT-related capabilities and their use of ICT for various class activities and teaching practices. They were also asked to indicate the frequency with which they used various ICT tools in their teaching of this class.
- *In your school:* The questions in this section asked the teachers about their views on using ICT in teaching and learning. The questions also asked teachers about provision for and practices concerning the use of ICT in their school.
- *Learning to use ICT in teaching:* This section asked teachers about whether their initial teacher education included learning to use ICT and whether they had participated in ICT-related professional learning.
- *Approaches to teaching:* This asked teachers to indicate the extent to which they agreed with a series of statements about using ICT in teaching and learning at school.

Development of the school principal and ICT coordinator questionnaires

The school questionnaires were designed to collect information about the school context in general and the use of ICT in teaching and learning in particular. Two questionnaires were used to collect this information. The first was directed to the school principal and the other to the school ICT coordinator. Both questionnaires were delivered online by default, but an alternative paper-based version was available in cases where respondents were unable or unwilling to complete it on a computer.

Factors relating to the school context included school characteristics, such as school size, management, and resources, the availability of ICT resources, professional development regarding ICT use for teachers, and expectations for ICT use and learning.

The questionnaire for school principals was designed to be completed in 15 minutes. The questions addressed school characteristics as well as school principals' perceptions of ICT use for teaching and learning at their schools.

The ICT coordinator questionnaire was also designed to be answered in 10 minutes. It included predominantly objective questions about the respective schools' ICT resources and their processes and policies with regard to this area.

The school principal questionnaire for the field trial included 18 questions with a total of 92 items, and was administered to 340 principals from the 14 countries that participated in the field trial. In most countries, about 24 school principals provided responses to the field trial questionnaire. The ICT coordinator questionnaire for the field trial consisted of 17 questions with a total of 91 items and was completed by 327 ICT coordinators at participating schools in 14 countries.

The analyses of field trial data focused on providing empirical evidence that would assist selection of the main survey material. However, the relatively small number of responses in each of the participating countries (the maximum was only one per school) meant that analyses of the field trial data gathered by the two questionnaires were limited in scope.

The ISC research team discussed the results of the school questionnaire field trial with NRCs before selecting the items that would be included in the final main survey instrument. Revisions made after the field trial included a rewording of some of the items.

The review of the field trial outcomes led to a reduction in the size of the school principal questionnaire, which consisted of 15 questions with a total of 94 items spread across the following four sections:

- *About you and your use of ICT:* This section asked school principals about their gender and ICT use.
- *Your school:* This section contained questions about school size, grades taught at the school, community size, and school management.
- *ICT and teaching in your school:* This section consisted of questions about the importance assigned to ICT use at school, monitoring of ICT use by teachers, and expectations about teacher use of ICT.
- *Management of ICT in your school:* This section contained questions about ICT management, ICT-related procedures, ICT-related professional development for teachers, and priorities for ICT use in teaching and learning.

The final ICT coordinator questionnaire comprised 15 questions (including two optional questions) with a total of 87 items (including 11 optional items). It contained the following three sections:

- *About your position:* This section asked ICT coordinators about their position at school and their school's experience with computers for teaching and learning.
- *Resources for ICT:* This second section included questions on the ICT equipment available at school.
- *ICT support:* This section consisted of questions on the support provided for ICT use at school and/or the extent to which a lack of resources was hindering that use.

Development and implementation of the national contexts survey

The ways in which students develop CIL and CT are potentially influenced by factors located at the country or national context level. These variables include, among others, the education system in general as well as policies on, and the curricular background of, CIL and CT education. The national contexts survey was designed to collect relevant data and information about both antecedents and processes at the country level. The experience of studies such as the Second Information Technology in Education Study (SITES) 2006 (Plomp et al. 2009), the US Department of Education (2011) study of educational technology, and ICILS 2013 (Fraillon et al. 2014) informed the development of the national contexts survey.

ICILS staff at the ISC at ACER organized the development and coordination of the national contexts survey as well as the analyses, verification, and reporting of the data collected by this instrument. Throughout this work, the ISC staff worked closely with national center staff from the participating countries.

The development and implementation work consisted of three phases:

- *Phase 1:* During this first phase, which spanned May to August 2017, the ISC team, in discussion with the national centers, reached agreement on the nature and scope of the survey's contexts and questions. During this phase, international project team members and national center staff discussed the various draft versions of the survey and reached agreement on a final version.
- *Phase 2:* Between March 2018 and January 2019, the NRCs answered the national contexts survey.
- *Phase 3:* The final phase took place between October 2018 and July 2019. During this phase, ISC staff reviewed the collected information and, where necessary, verified the outcomes with national centers.

During the development phase of the national contexts survey, the research team applied the following criteria when considering which contexts and questions to include in it:

- Relevance of content with regard to the ICILS 2018 assessment framework;
- Relevance and additional value of gathering information about the wider community context of CIL and CT education;
- Appropriateness of the instrument for the national contexts of the participating countries; and
- Validity of the measured variables in terms of comparability, analysis, and reporting.

The following issues were considered for modification, refinement, and further development of the ICILS 2013 national contexts survey:

- The increasing role of tablet-based education tools in education;
- Changes in policies related to ICT in education since 2013;
- Issues related to specific regions or groups of countries;
- Alternative response formats to improve the measurement of aspects of educational policies and curricula;
- Ways of increasing the objectivity of information provided by asking a greater proportion of factual questions and involving national experts to a greater extent; and
- Outcomes of a review reflecting the extent to which national contexts survey data from ICILS 2013 had proved useful for reporting.

The final version of the national contexts survey was placed, along with accompanying notes for guidance, online via servers at IEA. National centers were requested to draw on expertise in the field of ICT-related education in their countries when answering the survey.

The survey consisted of 25 questions including 163 items (some items were fixed responses and others asked for text response) plus 26 requests for an elaboration or comment. The questions asked respondents about key antecedents and processes in relation to CIL and CT education in their country. The questions were grouped into five sections:

- Education system;
- Plans and policies for using ICT in education;
- ICT and student learning at lower-secondary level;
- ICT and teacher development; and
- ICT-based learning and administrative management systems.

The online facility enabled national center staff to complete the survey in several administration sessions (i.e., they could log on and off in order to complete the questionnaire as needed information became available).

The ISC used the outcomes of the national contexts survey in conjunction with data from published sources to inform the descriptions of the education systems participating in ICILS. To ensure a maximum of accuracy regarding the information reported in the international report, national centers were invited to conduct detailed reviews of all reported outcomes that were based on the national contexts survey data collection.

References

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030193881

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*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783319142210.

Plomp, T., Anderson, R.E., Law, N., & Quale, A. (Eds.). (2009). *Cross national policies and practices on information and communication technology in education* (2nd ed.). Greenwich, CT: Information Age Publishing.

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *vol. 2*, 113–135.

US Department of Education, Office of Educational Technology. (2011). *International experiences with educational technology: Final report*. Washington, DC: Author.

CHAPTER 5:

Instrument preparation and verification of the ICILS 2018 study instruments

Sandra Dohr, Tim Friedman, David Ebbs, and Lauren Musu

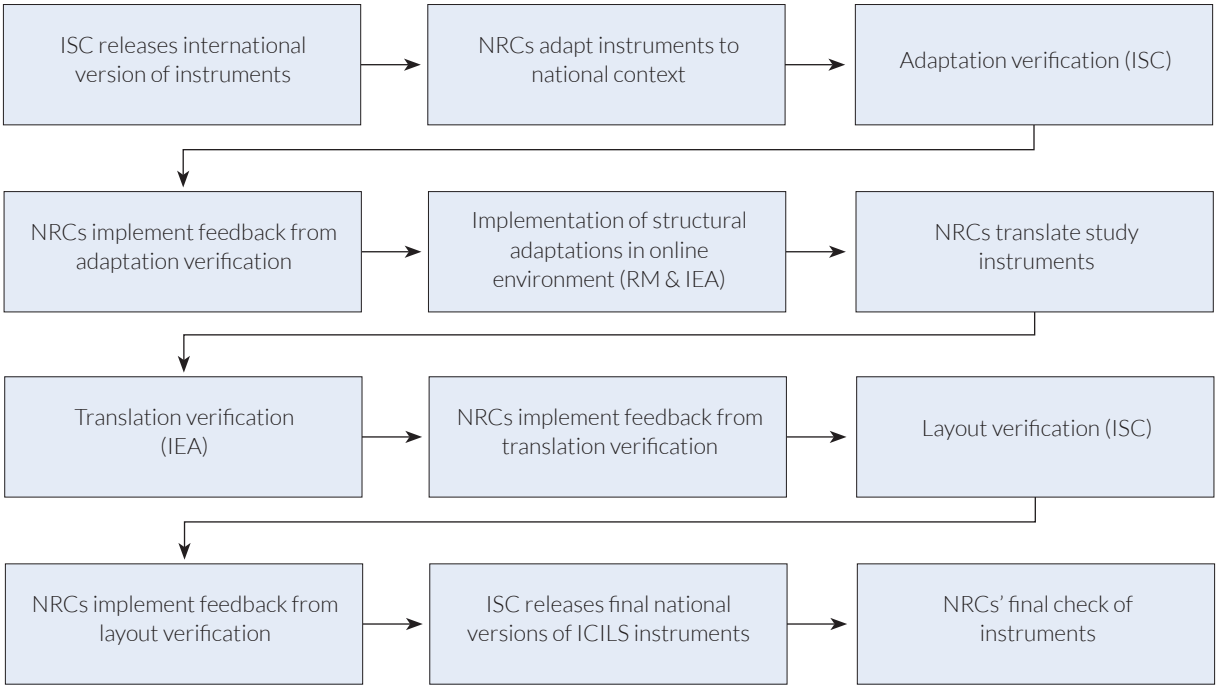
Introduction

The ICILS 2018 international study center (ISC) at ACER developed the English international version of the ICILS student assessment and context questionnaires in close collaboration with the national study centers of the ICILS participating countries and benchmarking participants. Translation, adaptation, and verification of these ICILS study instruments are part of the instrument preparation process and play a key role in ensuring high-quality data and international data comparability. All participating countries and benchmarking participants in ICILS 2018 were required to translate the international version of the ICILS study instruments and adapt them to their national context following detailed guidelines listed in the ICILS 2018 Survey Operation Procedures Unit 3.

The ultimate aim of the instrument preparation process is to guarantee comparability of the national instruments to the international version as well as equivalence across countries and languages but at the same time appropriateness for each country's national context and education system. In order to achieve the overarching goal of cross-national comparability and data validity, all national ICILS study instruments underwent a thorough and highly standardized preparation process that comprised several verification stages: adaptation verification, translation verification, and layout verification. The ISC at ACER and IEA oversaw the instrument preparation and verification processes and were responsible for different stages of the whole process. The ISC was responsible for reviewing and approving all national adaptations of the study instruments as well as the layout verification of the student instruments and paper-based questionnaires. IEA implemented national adaptations, affecting the structure of the principal, teacher, and ICT coordinator questionnaires and was responsible for preparing these questionnaires in the IEA Online Survey System, which was used for online administration of the questionnaires. IEA 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.

Figure 5.1 shows the overall workflow for preparing the national ICILS study instruments starting with the release of the international version of the ICILS study instruments, the adaptation and translation process, the three verification stages, and ending with the final set of study instruments ready for administration.

Figure 5.1: ICILS 2018 instrument preparation workflow



Notes: NRCs = national research coordinators; RM = RM Results (developed the software systems underpinning the computer-based student assessment instruments for the main survey).

Adaptation and translation of the ICILS study instruments

The adaptation and translation process

As part of the instrument preparation process, national research coordinators (NRCs) were required to translate the international source version of the ICILS study instruments, which included the assessment items as well as the questionnaires, and adapt them to their specific study needs and cultural context. The consortium provided national centers with the Survey Operation Procedures Unit 3, which described the procedures and guided NRCs through all necessary steps.

Documenting and implementing national adaptations

In ICILS 2018, there were two different types of adaptations that participants could apply to the instruments, namely structural and non-structural adaptations. Structural adaptations refer to alterations that change the structure of the study instruments. This includes adding national questions, adding, splitting, or omitting answering options or not administering questions. National centers were allowed to apply this type of adaptation solely to the background questionnaires and not to the assessment instruments.

Non-structural adaptations (also referred to as localizations) are defined as any adaptation which does not alter the structure of the study instruments, but may change the meaning to fit the local context. NRCs could add non-structural adaptations to their instruments, wherever they considered these necessary. Certain terminology cannot be translated directly in every target language and therefore needed to be replaced by terms or phrases that fit the national context and were appropriate for the target audience. Measurement and punctuation conventions are typical examples of such adaptations.

While NRCs could decide where they consider adaptations as crucial, in some cases they were required to adapt certain terminology. The international version of the study instruments indicated where such required adaptations were necessary. For example, required adaptations were the [target grade], personal names (e.g., [Male name 1]), places (e.g., [M-town]), or International Standard Classification of Education (ISCED) level equivalents. Fictional technical and software-related terminology also required adaptation in some cases. For instance, the term [WebSearch] referred to an online search engine and needed to be adapted to a term that makes sense to students.

In the field trial (where adaptations were requested directly into the translation system), structural and non-structural adaptation took place as two separate stages. However for the main survey, NRCs had to record all structural and non-structural adaptations for their study instruments at the same time in a document called the national adaptation form (NAF). During the adaptation and translation process, NRCs had to document the national version of the adapted text in the NAF and include an English back translation and an explanation or comment if necessary.

While national adaptations are important to reflect the country context, maintaining the comparability to the international source version is highly important. As such, the meaning and difficulty level must be preserved. Therefore, cognitive items may not be simplified or clarified in a way that would help the students identify the correct answer. In order to help countries adapting their materials to the national context, the ISC provided them with detailed notes that included definitions about specialized terminology in the ICT context as well as general guidelines for translation and adaptation.

Translating the ICILS 2018 study instruments

In order to create a high-quality translation, it is important to use professional and experienced translators and reviewers. For ICILS, IEA recommended using at least one translator and one translation reviewer for each target language, or more if only a limited amount of time was available. Ideally, the selected individuals should fulfill the following criteria:

- Translators and reviewers should have an excellent knowledge of English.
- Translators and reviewers should be native speakers of the target language.
- Translators should be familiar with survey instruments.
- Translators should have experience in translating electronic texts, such as websites.

The ideal workflow suggested that the selected translator translates all materials initially. In the next step, the reviewer reviews the translations, makes comments on the appropriateness of the translations and proposes improvements. The NRC finalizes the materials before submitting them for translation verification. In countries with more than one administered language, the same logic applied to each language. In addition, countries using more than one administered language needed to perform equivalence checks between the two or more language versions.

In countries where English was the administered language, the process of instrument preparation was identical to all other languages, except that the source version of the instruments did not require translation, but adaptation to accommodate the national English usage in their country.

Guidelines for translation and adaptation

Due to the complexity of languages and national contexts, it is difficult to provide participating countries and benchmarking participants with explicit guidelines for translation and adaptation. However, IEA provided NRCs with general guidelines and recommendations (included in the Survey Operation Procedures Unit 3), which are helpful to produce study instruments that are comparable to the international source.

During the translation process, translators, reviewers, and NRCs should pay particular attention to the following:

- Cultural and contextual differences should be identified and minimized.
- Terms and phrases in the target language must be equivalent to those in the international version.
- The essential meaning of the text and the reading level should be maintained.
- The translated text should reflect the same language level and degree of formality as the source text.
- The translated text should have correct grammar and usage.
- The translated text should not clarify or omit text from the source version, and should not add more information.
- The translated text should have equivalent qualifiers and modifiers appropriate for the target language.
- The translated text should have the equivalent social, political, or historical terminology appropriate for the target language and used at this level of education.
- Idiomatic expressions should be translated appropriately and not literally.
- Translators should avoid vocabulary, expressions, and concepts that may be unfamiliar to the target audience.
- Spelling, punctuation, and capitalization in the target text should be appropriate for the target language and the country's national context.
- National adaptations should be documented and implemented appropriately.
- Changes in layout due to translations should be reduced to a minimum.

Target languages

All participating countries and benchmarking participants in ICILS 2018 had to translate the student instruments into the relevant languages in their country. The selection of target languages usually includes all official languages of a country that are used in public areas of society or are the dominant languages of instruction. Minority languages are taken into account in some countries. The decision on whether a language was administered or not mainly depended on the sampling frame and whether or not minority schools were selected. In these cases, participating countries had to prepare their study instruments in all required languages. Most of the 14 participants administered only one language. However, two countries administered ICILS in two languages, namely Kazakhstan (Russian and Kazakh) and Finland (Finnish and Swedish). Luxembourg was the only country administering three languages (English, French, and German). In total, 17 different language versions were administered in ICILS 2018. Table 5.1 shows all administered languages in ICILS 2018.

ICILS instruments requiring translation and adaptation

The ICILS 2018 study instruments included two main components: student assessment and context questionnaires. The student assessment comprised seven different modules in which five were related to computer and information literacy (CIL) and two were related to computational thinking (CT). Three out of the five CIL modules were trend modules and therefore also administered in ICILS 2013. For countries that participated in both ICILS 2013 and ICILS 2018, these trend modules were used to compute trend measurement on the achievement scale. In order to ensure a valid trend measurement, the trend translations needed to remain identical to the previous cycle. The administration of the two CT modules (new to ICILS in 2018) was optional.

Table 5.1: Languages used for the ICILS 2018 study instruments

Participating country	Administered language
Chile	Spanish
Denmark	Danish
Finland	Finnish Swedish
France	French
Germany*	German
Italy	Italian
Kazakhstan	Kazakh Russian
Korea, Republic of	Korean
Luxembourg	English French German
Portugal	Portuguese
United States	English
Uruguay	Spanish
Benchmarking participants	
Moscow (Russian Federation)	Russian
North Rhine-Westphalia (Germany)*	German

Notes: *The same translations were used in Germany and North Rhine-Westphalia

In addition to the student assessment modules, NRCs had to prepare the following context questionnaires for data collection:

- Student questionnaire
- Teacher questionnaire
- Principal questionnaire
- ICT coordinator questionnaire

National study centers also had to translate and adapt the school coordinator and test administrator manuals as well as the scoring guides for constructed-response items. However, the ISC and IEA did not monitor the translation and adaptation process for these documents and did not perform any of the verification processes.

Translation systems

In order to prepare the ICILS 2018 study instruments for the main survey, the ISC and IEA provided national study centers with two online translation systems: the IEA Translation System and the AssessmentMaster (AM) system from RM Results, AM Designer (see Chapter 3 for more details about the two systems). The online translation systems served as the main platform for all involved parties to produce high-quality national instruments and to communicate with one another.

The instrument preparation process of the ICT coordinator, principal, and teacher questionnaire took place in the IEA Translation System, whereas AM Designer was used for the student assessment and the student questionnaire. Both systems allowed users to apply and edit their translations. The systems also kept a full record of the editing history for the text and enabled all users to retrace changes made by other users ensuring transparency of the translation process. In addition, the systems displayed the differences between a previously saved translation and the current translation in the form of track changes. This function supported NRCs with their reviewing process.

International verification processes

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

Adaptation verification

NRCs were asked to consult with ISC staff for a review of all proposed national adaptations. The ISC particularly emphasized the need for NRCs to discuss any adaptation that might result in a serious deviation from the international instruments. National centers began completing the NAF (Version I) after reviewing the international version of the survey instruments. They then submitted the NAF for consultation with the ISC, where two staff members reviewed the adaptations independently. Once completed, these reviews were consolidated into one document, and were then provided to the national centers with feedback on their adaptations and, where necessary, suggestions for better alignments to the international source version.

Common issues identified during review of adaptations included the following:

- 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 ISC asked the national centers to take the recommendations into account and update the forms accordingly. The NAF was only finalized once the ISC and national center were in agreement about all adaptations. Once this was done, these updated forms (Version II) would then inform the translation verification process.

Translation verification

Translation verification is the second out of three verification steps conducted during instrument preparation. The goal of this process is to ensure high-quality translations of the national ICILS 2018 instruments and their equivalence to the international source version. IEA managed and coordinated this process in cooperation with cApStAn Linguistic Quality Control.

International translation verifiers and their responsibilities

The contracted, professional, international translation verifiers were native speakers of the administered language. Further, the verifiers were certified translators working in English, with a university degree and ideally living in the target country or at least having experience working in the country context. Verifiers needed to attend a training seminar, where they received information regarding the ICILS study design and online environment. They also received instructions regarding the verification specifications and requirements.

The translation verifiers' main responsibilities included reviewing the translations for the target country and evaluating the accuracy and comparability of the national version of the ICILS instruments to the international version. Their tasks further comprised documenting all deviations in the country's translations and adaptations and suggesting alternatives to improve the quality of the translations. Verifiers checked if the meaning and reading level of the text had been affected and if the test items had been made easier or more difficult. They also ensured that no information was added or omitted. Translation verifiers made sure that the instruments contained all correct items and response options in the right order and that adaptations had been recorded and implemented correctly.

In addition to the general verification of the translations, international verifiers had to perform a trend check for countries that participated in ICILS 2013. To ensure the quality and accuracy of trend measurement on the ICILS achievement scale, it was of high importance to maintain the exact wording of the translation that countries used in the previous ICILS cycle. This was applicable to four countries that participated in ICILS 2018. IEA provided the verifiers with files that contained the final translations from ICILS 2013. The verifiers' task was to compare the translation the country submitted for translation verification with the version of the last cycle and flag any detected differences.

Translation verification procedure and verifiers' feedback

The translation verification process occurred in the two previously described online translation systems. International verifiers could apply corrections or suggestions directly to the text elements in the online environment. Any changes made by the verifier were displayed as track changes in the text element. In addition, verifiers were required to document their corrections and suggestions, describe the issue, and provide an English back translation in the form of comments for every affected text element and in the NAF. To simplify the verification process and to enhance the comprehensibility of the verifiers' feedback, verifiers assigned so-called severity codes to every change or suggestion. They made use of the following codes to indicate the severity of the issues they found:

- *Code 1: Major change or error*. The translation contained a severe error that affected the meaning or difficulty of the item. Examples included mistranslations, translations that change the meaning of a question, omitted or added information, incorrect order of questions or response options, and incorrectly implemented national adaptations. Verifiers also used this code to flag differences for trend translations.
- *Code 1?*: Used by verifiers whenever they were in doubt about the severity of an issue or unsure of how to correct a possible error.
- *Code 2: Minor change or error*. The translation contained a minor error that did not affect the comprehension. Examples included spelling errors, grammatical errors, and syntax errors that did not affect the comprehensibility of the question.
- *Code 3: Suggestion for alternatives*. Used by verifiers to suggest an alternative wording for an otherwise appropriate translation.
- *Code 4: Acceptable change*. The translation was deemed acceptable and appropriate. Also used by verifiers to indicate that a national adaptation had been documented and implemented correctly.

After translation verification, IEA reviewed the verifiers' feedback and returned the materials to the national centers. 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 procedure required NRCs to react to every comment the verifier made in the IEA Translation System, AM Designer, and the NAF. In principle, NRCs could accept, modify, or reject the verifier's suggestion. For the latter two options, NRCs had to give an explanation for why they changed or disagreed with the suggestion.

Some of the typical errors found by the translation verifiers during the verification work included mistranslations, literal translations, inconsistencies of terms or phrases, omissions or additions of text, undocumented adaptations, grammar and punctuation issues, and spelling mistakes. Some of the domain-specific concepts such as ICT-related technical terminology were a particular challenge to translate into some languages. According to the survey activities questionnaire completed by NRCs and used to collect feedback on survey operations, almost all participants found the translation verification feedback very useful (8) or at least somewhat useful (3). The majority of the national centers (10) reported that they did not experience any major problems during the

translation verification process. The constructive feedback aided NRCs in revising the materials and in improving the quality of their national versions in line with the translation guidelines for ICILS 2018.

As an additional quality control step, international quality control observers reviewed the implementation of the translation verification feedback after the data collection period and documented irregularities (see Chapter 9 for quality assurance procedures).

Layout verification of student instruments

Once translation verification had been completed, the ISC asked national centers to make any changes resulting from translation verification, and notify the ISC that their student materials were ready for layout verification. In AM Designer they were asked to change the status of their materials to "Ready for Layout Verification." The ISC accessed these materials and documented all issues in a worksheet added to the NAF. The layout issues in each set of instruments were grouped as to whether there were general layout issues relating to the set of instruments, or whether they related to a specific question or specific group of questions within an instrument.

The most common layout issues that were identified were:

- Missing or additional line breaks
- Unrealistic URLs or email addresses
- Text not fitting within the pre-defined spaces
- Issues with HTML tags
- Extra spaces
- Issues with structural changes to national content in the questionnaires

Technical issues relating to the translation software were often fixed by the verifiers on behalf of the national center or in some cases referred to the software developers.

National centers were provided with a summary of all layout issues. In cases where layout issues were considered minor, national centers were given feedback and were asked to make the appropriate changes to their materials without the need for further verification. In cases where more substantial layout issues were identified, national centers were provided with detailed feedback concerning all issues and were asked to resubmit their materials for further layout verification.

Layout verification of teacher, principal, and ICT coordinator questionnaires

All countries and benchmarking participants administered the teacher, principal, and ICT coordinator questionnaires online. Once the content for these questionnaires was finalized, countries were asked to notify IEA who prepared each questionnaire accordingly by importing the translations from the IEA Translation System into the IEA Online Survey System. In case of unclear or unresolved translation issues following translation verification, IEA asked the countries for clarification. Otherwise, all layout issues identified were fixed by IEA. Countries were then provided with the questionnaires for review and could submit any request for changes to IEA to implement.

Some countries also opted to implement some questionnaires using paper-based form. Countries were provided with paper questionnaires which IEA extracted from the IEA Translation System. Prior to sending the paper questionnaires to the countries and to ensure that data from both administration modes were comparable, IEA conducted a systematic check of the paper and online questionnaires. Apart from a few inevitable exceptions any deviations with regard to content and layout between paper and online instruments were reported back to the countries.

Before the national questionnaires were published online, the layout and structure of all online questionnaires were checked thoroughly by IEA. Visual checks were run using the same standards and procedures as for the verification of the paper layout. In addition, the structure of the national online instruments was checked against the structure of the international online instruments (e.g., number of categories or width of non-categorical questions). Only intended deviations which were documented in the NAF were approved.

Once all identified inconsistencies had been fixed, the questionnaires were published online. At this stage the country-specific URLs were activated and set to "review mode," which is a built-in safeguard within the IEA Online Survey System that prevents actual respondents from logging into online questionnaires before finalization. Without this safeguard, users would be able to access the questionnaires and view the welcome page. Prior to publishing the online questionnaires, i.e., starting to collect actual respondent data, the questionnaires underwent the subsequent two review steps:

(1) The ISC conducted a layout check. All inconsistencies were documented and sent to IEA for implementation.

(2) As a last check, the NRCs were asked to carry out a final review of the questionnaires in the online environment. In a few cases, this resulted in additional minor changes (e.g., correction of spelling errors).

Only after the finalization of the online questionnaires were respondents notified and provided the link and login information to access the questionnaires.

Summary

To ensure high-quality translations and comparability of the national instruments prepared by national centers to the international instruments, the 2018 cycle of ICILS incorporated stringent procedures for adaptation, translation, and international verification, similar to the previous cycle. NRCs were provided with a comprehensive set of guidelines and manuals that facilitated the production of national instruments. In each country, the preparation of national instruments started with the adaptation and translation of the international version of the ICILS 2018 materials into the identified language(s) of administration. Upon completion of the translations and adaptations, the national instruments first underwent international adaptation verification and then international translation verification, which involved a thorough review of the translated materials by external linguistic experts. After these two steps, the layout of all national materials was verified. Taken together, these stages of national instrument production ensured that the adaptations, translations, and layout of all national study instruments underwent thorough verification.

CHAPTER 6:

Sampling design and implementation

Sabine Tieck

Introduction

This chapter provides an introduction to the sampling design and the implementation of the ICILS 2018 student, teacher, and school survey which are consistent with those used in ICILS 2013 (see Meinck 2015). International comparative surveys require samples to be drawn randomly to guarantee valid inferences from the observed (sampled) data on the population features under study. Randomness in the selection of study participants is a key criterion to ensure international comparability or comparability between subnational entities. ICILS determined an international survey design that considered all requirements for sampling quality specified in the *Technical Standards for IEA Studies* (Martin et al. 1999).

The ICILS 2018 sample design is referred to as a "complex" design because it involves multi-stage sampling, stratification, and cluster sampling. This chapter describes these features and provides details on target population definitions, design features, sample sizes, and achieved design efficiency. It focuses on presenting the international standard sampling design whereas specific characteristics of each national sampling plan are given in Appendix B.

The samples of schools, students, and teachers within each ICILS country and education system were selected or verified by IEA in collaboration with the national research coordinators (NRCs) of each participating country or educational system. A series of manuals and the IEA Windows Within-School Sampling Software (IEA WinW3S) supported NRCs in their sampling activities. The sampling referee Marc Joncas gave advice on sampling methodology, as well as reviewed and adjudicated all national samples for this study.

Target population definitions

When conducting a cross-country comparative survey, it is important to clearly define the target population(s) under study. ICILS collected information from students, their teachers, and their schools, which required clear definitions for all three populations. The definitions enabled ICILS NRCs to correctly identify and list the targeted schools, students, and teachers from which samples were to be selected.

Definition: Students

ICILS defined the target population of students as follows:

The student target population in ICILS consists of all students enroled in the grade that represents eight years of schooling, counting from the first year of ISCED Level 1,[1] *providing the mean age at the time of testing is at least 13.5 years.*

For most countries, the target grade was the eighth grade or its national equivalent. Italy followed the international population definition but decided to survey their grade 8 students (and their teachers) at the beginning of the school year. Due to this deviation, their results were annotated accordingly in the international report. 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.

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

1 ISCED stands for International Standard Classification of Education (UNESCO 1997).

Definition: Teachers

ICILS 2018 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 that fulfill the following conditions: They are 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 since the beginning of the school year.[2]

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 of the curriculum (e.g., cases where religion is not a regular subject and taught by external persons); and
- Teachers who have joined a school after the official start of the school year.

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

Definition: Schools

In ICILS 2018 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 enroled.

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:

- Sub-units of larger "campus school" (administrative "schools" consisting of smaller schools from different cities or regions) should be regarded as separate schools for sampling purposes. If a part of a larger campus school was selected for ICILS 2018, the principal or ICT coordinator of the combined school was asked to complete the school questionnaire with respect to the sampled sub-unit only.
- Schools consisting 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 for sampling purposes.
- The parts of a school with two or more different study programs that have different teaching staff, take place in different buildings, and offer no opportunity for students to change from one study program to the other, should be regarded as two or more separate schools for sampling purposes. The study programs should be listed as separate units on the school sampling frame.

2 Teachers that are on a long-term leave during the testing period (e.g., maternity or sabbatical leave) are not in scope of ICILS.

Coverage and exclusions

Population coverage

The ICILS consortium encouraged participating countries to include all schools, students, and teachers defined in the target populations in the study, in order to ensure a full coverage of these target populations.

However, it was deemed appropriate to exclude some schools, students, and teachers from the target population for practical reasons, such as difficult test conditions or prohibitive survey costs. Some students and teachers were not surveyed due the removal of their entire school from the sampling frame (*school-level exclusions*) while some students (but not teachers) were excluded within participating schools (*within-sample exclusions*).

As in other large-scale assessments, it should be emphasized that the ICILS 2018 samples represent the nationally defined target populations only (without the excluded members of the internationally defined target population).

School-level exclusions

Table 6.1 gives an overview of the types of exclusions of schools and their respective percentages of all schools in the desired national target population within each participating country. The (school-level) percentages ER_{sch} were computed as:

$$ER_{sch} = \frac{ER_{sch}}{TP_{sch}} \text{ x100}$$

ER_{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.

In most countries, very small schools and schools exclusively dedicated to students with special needs students were excluded. Frequently, schools following a curriculum that differed from the mainstream curriculum were also not part of the nationally defined target population. Because school-level data (collected via the principal and ICT coordinator questionnaires) in the ICILS 2018 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 differed significantly across countries, a point that should be kept in mind when interpreting results from school-level data. Please note also that because school exclusions typically concern small schools, the percentages of excluded schools always tend to be higher than the corresponding percentages of excluded students or teachers.

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

Country	Type of exclusion	Excluded schools (% of all schools)
Chile	Very small schools (fewer than six students)	5.0
	Special needs schools	0.1
	Geographically inaccessible	0.1
	Total	5.1
Denmark	Very small schools (fewer than five students)	7.0
	Special needs schools	5.1
	Treatment centers	1.6
	German, English, Waldorfs schools	1.3
	Total	14.9
Finland	Special needs schools	9.5
	Language schools (instructional language not Finnish or Swedish)	1.0
	Total	10.5
France	Overseas territories (TOM)	1.4
	Mayotte	0.3
	Private schools without contract	8.3
	Specialized schools	0.8
	Total	10.8
Germany	Special needs schools	8.9
	Very small schools (fewer than three students)	1.4
	Total	10.4
Italy	Very small schools (fewer than six students)	0.2
	Special needs schools	0.1
	Students taught in Slovene	0.1
	Schools in remote geographical area or in little islands	0.1
	Total	0.5
Kazakhstan	Students are taught in Uzbek language	1.2
	Students are taught in Uighur language	0.2
	Students are taught in Tadjik language	0.1
	Students are taught in other language	0.0
	Special needs schools	1.3
	Very small schools (fewer than four students)	5.8
	Total	8.6
Korea, Republic of	Very small schools (fewer than five students)	1.8
	Geographically inaccessible schools	4.4
	Physical education school	0.3
	Total	6.6
Luxembourg	No exclusions on school level	0.0
Portugal	Very small schools (fewer than seven students)	2.0
	International schools	1.1
	Total	3.1
United States	No exclusions on school level	0.0

Table 6.1: Percentages of schools excluded from the ICILS target population (contd.)

Country	Type of exclusion	Excluded schools (% of all schools)
Uruguay	Special needs schools	0.2
	Geographically remote schools	8.8
	Total	9.0
Benchmarking participants		
Moscow (Russian Federation)	Special needs schools	2.5
	Very small schools (fewer than seven students)	4.8
	Total	7.3
North Rhine-Westphalia (Germany)	Special needs schools	9.7
	Very small schools (fewer than three students)	0.2
	Total	9.8

Student-level exclusions

Each country was required to keep the overall rate of excluded students (due to school-level and within-school exclusions) below five percent (after rounding) of the desired target population. In three education systems participating in ICILS 2018 the overall exclusion rate was above five percent, which resulted in respective annotations in the ICILS 2018 international report (Fraillon et al. 2020). Table 6.1 and Appendix B of this report provide details about the exclusion types for each country.

The overall exclusion rate of students is the sum of the students' school-level exclusion rate and the weighted within-sample exclusion rate. Table 6.2 provides the respective percentages for ICILS 2018 countries.

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

Country	Students' school-level exclusion (%)	Within-sample exclusions (%)	Overall exclusions (%)
Chile	0.5	0.8	1.3
Denmark	3.1	4.4	7.5
Finland	1.6	2.4	4.0
France	3.4	1.3	4.7
Germany	1.5	2.9	4.3
Italy	0.1	2.9	3.0
Kazakhstan	3.4	2.1	5.6
Korea, Republic of	0.9	0.6	1.5
Luxembourg	0.0	3.9	3.9
Portugal	0.8	8.0	8.9
United States	0.0	5.0	5.0
Uruguay	1.1	0.0	1.1
Benchmarking participants			
Moscow (Russian Federation)	0.7	2.3	3.0
North Rhine-Westphalia (Germany)	1.4	3.1	4.6

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:

$$ER_1 = \frac{E_1}{TP} \text{ x100}$$

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 with physical or mental disabilities or students who could not speak the language of the test (usually, students with less than one year of instruction in the test language). Students could be excluded prior to the within-school sampling or after the within-school sampling was performed.[3] The percentage of student within-school exclusions was calculated using the number of students excluded within schools and the total number of students belonging to the national desired target population.

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

$$ER_2 = \frac{W^{E}_2}{W^{P} + W^{E}_2} \text{ x } (1 - ER_1) \text{ x } 100$$

W^{E}_2 is the sum of weights of excluded students 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. The students' school-level exclusion rate is taken into account by multiplying 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:

$$ER_{tot} = ER_1 + ER_2$$

Table 6.2 provides the respective percentages for ICILS 2018 countries.

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 about the total number of teachers teaching in the target grade as well as the proportion of teachers teaching in the target grade in excluded schools. For Germany, Finland, and North Rhine-Westphalia (Germany), no statistics on the number of eligible ICILS 2018 teachers were available and therefore it was not possible to compute exclusion rates. Teacher exclusion rates exceeded five percent in Denmark, France, and the United States.

3 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 enroled students within the exclusion categories, which resulted in the corresponding school(s) to be excluded ex post.

School sampling design

IEA used a stratified two-stage probability cluster sampling design in order to conduct the school sample selection for all ICILS 2018 countries. During the first stage, schools were selected systematically with probabilities proportional to their size (PPS) as measured by the total number of enroled target grade students. During the second stage, within participating schools, students enroled in the target grade were selected using a systematic simple random sample approach.

The following subsections provide further details on the sample design for ICILS 2018.

School sampling frame

In order to prepare the selection of school samples, national centers provided a comprehensive list of schools including the numbers of students enroled in the target grade. This list is referred to as the *school sampling frame*. To ensure that each ICILS 2018 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 eligible school in the sampling frame:

- A unique identifier, such as a national identification number;
- School's measure of size (MOS), which was usually the number of students enroled in the target grade or an adjacent grade; and
- Values for each of the intended stratification variables.

Stratification of schools

Stratification is part of many sampling designs and entails the grouping of sampling frame units by common characteristics. Examples for such groups of units (schools in the case of ICILS 2018) would be geographic region, urbanization level, source of funding, or performance level.

ICILS 2018 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.

Generally, IEA studies use stratification for the following reasons:

- They use implicit or explicit stratification to improve the efficiency of the sample design, thereby making survey estimates more reliable and reducing standard errors (to this end national centers identify stratification variables expected to be closely associated with students' learning-outcome variables).
- They use explicit stratification to apply disproportionate sample allocations to specific groups of schools.

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% of students in private and 90% 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 after discussions with the 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)
Chile	Grade (Grade 8 and 9/Grade 8 only) (2) Administration (public/private/private subsidized) (3) Urbanization level (2)	10	Performance level (4)
Denmark	None	1	National achievement score (5)
Finland	Language of instruction (2) Region (Helsinki & Uusimaa/Southern/Western/Northern & Eastern) (4) Urbanization (2) Within Swedish speaking school: Region (2)	9	Within Western, and Northern & Eastern stratum: Region (4) Within Swedish speaking strata: Urbanization (2)
France	School administration (3) Urbanization (3) Digital equipment level (2)	18	None
Germany	North Rhine-Westphalia/ other federal states (2) Track (gymnasium/nongymnasium/special needs schools) (3)	5	SES indicator (3) Federal state (16)
Italy	Region (North/Central/South) (3)	3	Administration (2) Performance level (5)
Kazakhstan	Urbanization (urban/rural) (2) Language of instruction (4)	8	None
Korea, Republic of	Urbanization (3) School gender (3)	9	None
Luxembourg	Schools following national curriculum/ Schools following other curriculum (2)	2	None
Portugal	Administration (2) Region (25)	28	None
United States	Poverty level (2) Administration (2) Region (4)	12	Urbanization (5) Ethnicity status (2)
Uruguay	Administration (2) Region (2) School type (2)	6	None
Benchmarking participants			
Moscow (Russian Federation)	Performance level (5)	5	Administration (2)
North Rhine-Westphalia (Germany)	Track (gymnasium/nongymnasium/special needs schools) (3)	3	SES indicator (3)

School sample selection

In order to select the school samples for the ICILS 2018 main survey, the sampling team used *stratified PPS systematic sampling*. This method is customary in most large-scale surveys in education, and notably in 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 probablility of selection. In reality, however, no single ICILS sample was self-weighting.

First, note that a sampling design can only be self-weighting for one target population, and ICILS aimed for 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 are not self-weighting because explicit stratification was used and because sampling allocation can hardly ever be exactly proportional for different explicit strata. Finally, samples of four out of 14 countries were disproportionally allocated to explicit strata in order to achieve precise estimates for subgroups.[4]

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 which case 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 within each implicit stratum by MOS (alternately sorted in increasing and decreasing order).
3. Calculating a sampling interval by dividing the total MOS by the number of schools to be sampled from that explicit stratum.
4. Determining a random starting point, a step that determines the first sampled school.
5. Selecting all following schools by adding the sampling interval to the random start and then subsequently to each new value every time a school was selected. Whenever the cumulated MOS was equal or above the value for selection, the corresponding school was included in the sample.

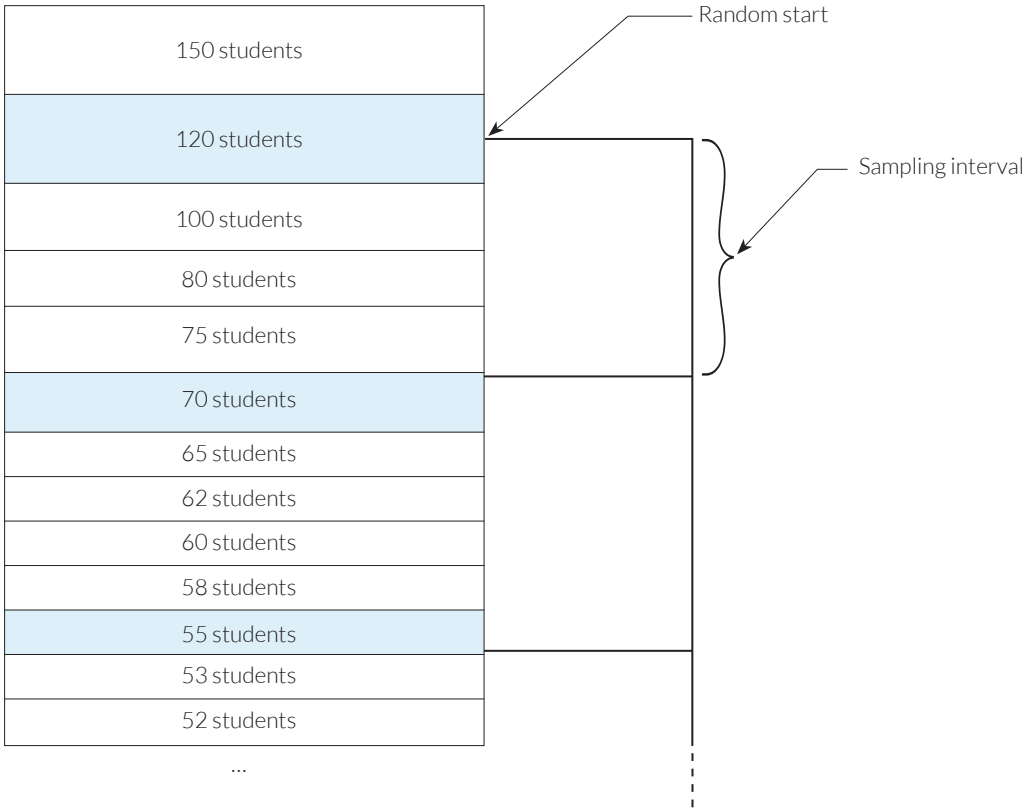
Figure 6.1 visualizes the process of systematic PPS sampling within an explicit stratum. In this diagram, the schools in the sampling frame are sorted descending by MOS, and the height of the cells reflects the number of target-grade students in each school. A random start determines the second school 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.[5]

4 See Chapter 7 for more details on sampling weights.

5 The corresponding information can be accessed in the file "TIMSS and PIRLS Sampling Schools" in "Sample Design Details" at https://timssandpirls.bc.edu/methods/pdf/Sampling_Schools.pdf

Figure 6.1: Visualization of PPS systematic sampling



Source: Zuehlke 2011.

In some 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 changing size measures. Usually a school was regarded as "small" if the number of enroled target grade students was less than 20.
- Very large schools (i.e., schools with more students than the value of the sampling interval) were placed into a separate 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 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 to agree 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 we prevented this by selecting the sample of schools to participate in each of the field trial and main survey simultaniously (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 2018 was conducted in the same year as the OECD's Teaching and Learning International Survey (TALIS) 2018 and Programme for International Student Assessment (PISA) 2018, and some national education surveys (e.g., the educational trends in Germany). Several countries requested that schools should be selected for one of these studies only to reduce the administrational burden for the schools. The ICILS 2018 sampling team collaborated closely with the staff implementing sampling for TALIS 2018 (at Statistics Canada) and for PISA 2018 (at Westat) 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.

Once schools had been selected from the sampling frame, up to two (non-sampled) replacement schools were assigned for each originally sampled school. The use of replacement schools in ICILS 2018 was limited (see Chapter 7 for more details). In order to reduce the risk of non-response bias due to replacement, the replacement schools were usually assigned as follows: The school, which 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 they belonged to, were of similar size, and belonged to the same stratum.[6]

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 IEA WinW3S software, developed by IEA, ensured the random selection of students and teachers within the sampled schools. Replacement of non-responding individuals was not permitted in ICILS 2018.

Student sampling

IEA WinW3S employed systematic stratified sampling with equal selection probabilities to select students from comprehensive lists of target grade students provided by the participating schools. Implicit stratificiation was applied to ensure a nearly proportional allocation among subgroups and to increase sample precision. Thus students were sorted by gender, class allocation, and birth year prior to sampling.

Teacher sampling

As was the case for student sampling, IEA WinW3S 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 implicilty stratified by sorting them according to gender, main subject domain, and birth year prior to sampling.

6 For very large schools (see above) it is not always possible to assign two replacement schools as the preceeding school or the school directly listed after the sampled schools is either another sampled school or a replacement school assigned to another sampled school. In extreme cases, no replacement school can be assigned. This could also happen if the number of schools to sample from a stratum is very high compared to the number of schools in this stratum. If the sampled school is the first or last in its stratum, usually the two following and preceeding schools are assigned respectively. Please note further that for the field trial only one replacement school was assigned.

Sample size requirements

The ICILS 2018 consortium, in line with 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 should ensure a specified level of precision for population estimates; defined by confidence intervals of ±0.1 standard deviation for means, and ±5% for percentages. With respect to the main outcome variable in ICILS 2018–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 one participating country (Unites States), which failed to meet the IEA sample participation standards, all participating countries and educational systems achieved this requirement (see Fraillon et al. 2020, p. 75). The required precision levels of percentages were also met for the vast majority of population estimates presented in the ICILS international report (Frallion et al. 2020).

There were also other consideration 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, Meinck and Vandenplas 2012);
- 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.[7] 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 of 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 schools with less than 20 students in the target grade was relatively large so that 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

Typically 20 students were randomly selected from the full target grade cohort (i.e., across all classes) in each sampled school. In schools with 25 or fewer students in the target grade, all students were selected.[8]

7 Luxembourg conducted a census of schools, i.e., all 41 school were asked for participation.

8 In the Unites States a minimum of 30 students were randomly selected, because students to be excluded could only be identified after the within-school sample was conducted. In Luxembourg, a census of students was used. Thus all students were selected.

Each country was required to have an achieved student sample size of about 3000 tested students. Due to non-response, school closures, or other factors, some countries did not meet this requirement. The ICILS 2018 sampling team did not regard this outcome as problematic as long as the country met the overall participation rate requirements (see Chapter 7).

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.[9]

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

- Schools: 150 in each country
- Students: 20 (or all) per school
- Teachers: 15 (or all) per school

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 differences in the numbers of participating schools for the student and teacher survey across ICILS 2018 countries.[10]

Table 6.4: School, student, and teacher sample sizes

Country	Originally sampled schools	Student survey		Teacher survey	
		Participating schools	Participating students	Participating schools	Participating teachers
Chile	180	178	3092	174	1686
Denmark	150	143	2404	138	1118
Finland	150	144	2546	143	1853
France	156	156	2940	122	1462
Germany	234	209	3655	182	2328
Italy	150	150	2810	148	1775
Kazakhstan	186	183	3371	184	2623
Korea, Republic of	150	150	2875	147	2127
Luxembourg	41	38	5401	28	494
Portugal	220	200	3221	208	2823
United States	352	263	6790	259	3218
Uruguay	177	166	2613	121	1320
Benchmarking participants					
Moscow (Russian Federation)	150	150	2852	150	2235
North Rhine-Westphalia (Germany)	115	109	1991	107	1468

9 In Luxembourg, the minimum sample size was increased to 25 teachers per school, due to the small number of schools.
10 Please refer to Chapter 7 for details on ICILS 2018 standards for sampling participation.

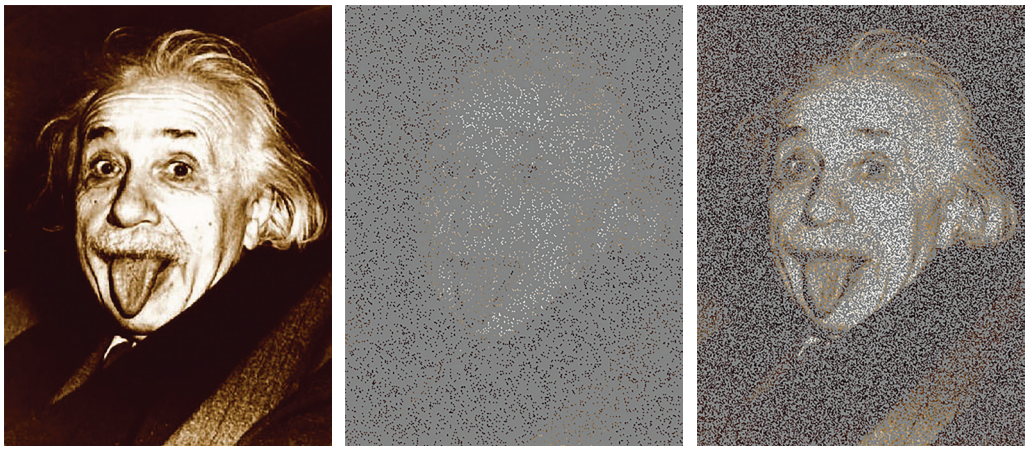
Efficiency of the ICILS 2018 sample design

As already noted, ICILS 2018 determined specific goals in terms of sampling precision, especially that standard errors should be kept below specific thresholds. Let us illustrate this concept in some more detail for readers who are not as familiar with this topic.

In any sample survey, researchers would like to use data collected from the sample to get a good (or "precise") picture of the population from which the sample was drawn. However, there is a need to define what is "good" in terms of sampling precision. Statisticians aim for a sample that has as little variance and bias as possible for specific design and cost limits. A measure of the precision is the standard error. The larger the standard error, the more "blurred" the picture is, and inferences from sample data to populations become less reliable.

Let us assume our population of interest is the left-hand picture in Figure 6.2 below, the famous picture of Einstein taken by Arthur Sasse in 1951. The picture consists of 340,000 pixels. We can draw samples with increasing numbers of pixels from this picture and reassemble the picture using only the sampled pixels. As can be seen in the middle and right-hand pictures in Figure 6.2, the picture obtained from the sampled pixels becomes more precise as the sample size increases. The standard errors from different samples sizes are equivalent to reflections of the sampling precision in this example.

Figure 6.2: Illustration of sampling precision—simple random sampling



Picture = Population | Sample size = 10,000 | Sample size = 50,000

Determining sampling precision in infinite populations is relatively straightforward as long as simple random sampling (SRS) is employed. The standard error of the estimate of the mean μ from a simple random sample can be estimated as:

$$\sigma_{\hat{\mu}} = \sqrt{\frac{\sigma^2}{n}}$$

with σ^2 being the (unknown) variance in the population and n being the sample size. If the variance in the population is known, the sample size needed for a given precision level can be easily derived from the formula. For example, assuming the standard deviation σ of an achievement scale to be 100, the population variance σ^2 would be 10,000, and the standard error of the estimated scale mean $\sigma_{\hat{\mu}}$ will equal five scale score points or less. Rearranging the formula above leads then to a required minimum sample size of 400 students per country. As pointed out earlier, however, the actual minimum sample size for participating countries in ICILS 2018 was 3000 students.

The key reason for this sample size requirement is that ICILS 2018 did not employ SRS sampling but cluster sampling. Students in the sample are members of "clusters" as groups of them belong to the same schools.

Students within a school tend to be more similar to one another than students from different schools because they are exposed to the same environment and teachers. Furthermore, they also often share common socioeconomic backgrounds. Therefore, the gain in information through sampling additional individual students within schools is less than when sampling additional schools, even if the total sample size is kept constant. In other words, due to the homogeneity of students in the same schools the sampling precision of cluster samples with similar sample sizes tends to be less than when applying SRS.

For this reason, the SRS formula given above is not applicable for data from cluster samples. In fact, and depending also on the outcome variable beeing measured, applying this formula will most likely underestimate standard errors from cluster sample data by a considerable margin. Figure 6.3 visualizes this effect through the example of the Einstein portrait where the number of pixels sampled is the same in both pictures. However, in the right-hand picture, clusters of pixels were sampled rather than single pixels as in the left-hand picture.

Figure 6.3: Sampling precision with equal sample sizes—simple random sampling versus cluster sampling

Note that stratification also has an influence on sampling precision. Through the choice of stratification variables related to the outcome variables it is possible to increase the sampling precision compared to non-stratified samples. However, experience shows that in large-scale assessments in education the impact of stratification on sampling precision tends to be much smaller than the effect of clustering. Stratification is another reason as to why the SRS formula for estimating sampling variance is not applicable for ICILS 2018 survey data.

Because of the above reasons, estimation of sampling variance for complex sample data is not as straightforward as it is for simple random samples. Chapter 13 of this report explains in more detail the jackknife repeated replication (JRR) method which should be used for a correct estimation of standard errors for ICILS 2018 data.

The achieved efficiency of the ICILS sampling design is measured by the design effect as:

$$deff = \frac{Var_{JRR}}{Var_{SRS}}$$

where VAR_{JRR} is the design-based sampling variance for a statistic estimated by the JRR method, and VAR_{SRS} is the estimated sampling variance for the same statistic on the same data base but

considering the sample as a simple random sample (with replacement, conditional on the achieved sample size of the variable of interest).[11] If we can estimate the design effect in a given country from previous surveys with the same or at least equivalent outcomes variables, we can also determine a desired sample size for a cluster sample design.

ICILS 2018 required an "effective" sample size of 400 students. Within the context of large-scale studies of education, the "effective" sample size is an estimate of the sample size that would be needed to achieve the same sampling precision of a cluster sample if simple random sampling had been applied. So, for example, in a country where the design effect in a previous survey was estimated at eight, multiplying this number by the effective sample size would provide an estimate of the desired sample size for the next survey, assuming that the samples apply the same design for stratification and clustering.

Table 6.5 to Table 6.7 provide the design effects of the ICILS 2018 main outcome variables for students and for teachers in each participating country. This information helps to determine sample sizes and design strategies in future surveys with similar objectives. The design effects vary for different scales but this is not unusual, since the similarity of students and teachers within schools should be higher when asking about school-related matters rather than, for example, about how frequently they use ICT as individuals. The last column of Table 6.6 and Table 6.7 provides estimates of the effective sample sizes.

In Table 6.5, we can see that the effective sample size relates to the design effect of the CIL and the CT scale, while in Table 6.6 the effective sample size relates to the average design effects across the presented scales. As is evident from Table 6.5, all national samples achieved or, more often, significantly exceeded the envisaged effective sample size of 400. Table 6.6 shows that the effective teacher sample size was also estimated at 400 or above in all ICILS 2018 countries.

The average design effect of the CIL scale is equal to 2.6. Most other scales pertaining to the student survey showed even lower design effects. The average design effect of the CT scale is 2.0 and tends to be lower than the CIL scale design effect in most ICILS 2018 countries that administered the optional CT assessment. The teacher-related scales had an average design effect of 2.6 across ICILS 2018 countries.

11 The measurement error for the CIL scales is included in VAR_{JRR}. Chapter 13 provides further details on measurement error estimation.

Table 6.5: Design effects of main outcome variables—Student survey

Country	Sample size	Design effects using scale scores pertaining to students' use of and engagement with ICT at home and school																
		S_SPECEFF	S_GENEFF	S_CODLRN	S_ICTLRN	S_ICTFUT	S_ICTNEG	S_ICTPOS	S_SPECLASS	S_GECLASS	S_SPECACT	S_GENACT	S_EXCOMP	S_EXSMART	S_EXTAB	S_USECOM	S_USEINF	S_USESTD	S_ACCONT
Chile	3092	2.1	3.4	2.9	3.1	2.7	1.6	2.4	2.9	3.6	3.9	3.4	2.4	3.1	2.2	2.7	3.0	2.0	1.9
Denmark	2404	1.6	1.2	1.3	1.5	1.4	1.4	1.2	2.1	2.3	1.6	2.5	1.6	1.0	1.0	1.3	1.9	2.0	1.8
Finland	2546	1.5	1.9	2.4	1.6	1.7	1.3	1.3	1.5	2.7	1.1	1.9	1.7	1.0	1.6	1.5	1.1	1.9	1.3
France	2940	1.0	1.6	1.8	1.6	1.1	1.5	1.9	1.6	2.0	1.5	1.2	1.5	1.3	2.2	1.9	1.4	1.7	2.1
Germany	3655	1.7	2.0	2.2	2.8	1.8	2.1	2.3	2.3	3.1	1.9	2.0	2.2	2.7	1.5	3.0	2.2	2.8	3.1
Italy	2810	1.6	1.6	2.3	2.1	1.3	1.4	1.3	1.8	2.4	1.6	2.0	1.4	1.6	1.2	1.0	1.3	1.9	1.0
Kazakhstan	3371	2.1	2.5	3.8	2.9	2.3	1.4	2.5	5.4	6.4	3.0	4.2	3.5	1.8	2.9	4.8	4.4	5.6	5.6
Korea, Republic of	2875	1.5	2.0	2.5	1.9	2.0	1.3	2.4	2.0	3.2	2.4	4.8	1.4	1.6	2.2	1.5	2.0	2.9	1.4
Luxembourg	5401	0.6	0.8	0.6	0.8	1.3	1.6	1.0	0.6	0.9	0.7	0.8	0.7	0.7	0.6	0.6	1.0	0.8	0.8
Portugal	3221	2.5	2.1	2.2	2.7	1.3	1.6	2.1	2.3	3.1	1.8	2.1	1.9	1.6	1.9	2.1	2.0	2.9	2.1
United States	6790	2.4	1.9	1.8	2.9	1.9	1.0	2.2	2.8	3.7	1.7	4.5	2.0	1.5	1.7	1.3	1.8	4.1	1.6
Uruguay	2613	1.9	1.6	1.8	1.6	1.0	1.4	1.5	1.7	2.6	1.5	1.1	1.5	1.2	1.6	2.4	2.1	1.3	2.1
Benchmarking participants																			
Moscow (Russian Federation)	2852	2.3	1.0	2.4	1.4	2.1	1.6	1.8	2.6	3.6	1.4	2.5	2.0	1.0	1.6	1.1	1.3	2.4	2.3
North Rhine-Westphalia (Germany)	1991	1.0	1.4	1.1	1.6	1.0	1.1	1.5	1.7	2.9	1.3	1.3	2.1	0.9	0.9	1.3	1.8	1.3	1.0
ICILS 2018 average	3326	1.7	1.8	2.1	2.0	1.6	1.5	1.8	2.2	3.0	1.8	2.4	1.8	1.5	1.6	1.9	2.0	2.4	2.0

Scale scores

S_SPECEFF ICT self-efficacy regarding the use of specialist applications
S_GENEFF ICT self-efficacy regarding the use of general applications
S_CODLRN Learning of ICT coding tasks at school
S_ICTLRN Learning of ICT tasks at school
S_ICTFUT Expectations of future ICT use for work and study
S_ICTNEG Negative perceptions of ICT for society
S_ICTPOS Positive perceptions of ICT for society
S_SPECLASS Use of specialist applications in class
S_GENCLASS Use of general applications in class
S_SPECACT Use of specialist applications for activities
S_GENACT Use of general applications for activities
S_EXCOMP Computer experience in years
S_EXSMART Smartphone experience in years
S_EXTAB Tablet experience in years
S_USECOM Use of ICT for social communication
S_USEINF Use of ICT for exchanging information
S_USESTD- Use of ICT for study purposes
S_ACCONT Use of ICT for accessing content from the internet

Table 6.6: Design effects and effective samples sizes of mean of scale scores and plausible values—Student survey

Country	Sample size	Design effects using:			Effective sample size based on:		
		Mean of scale scores	CIL plausible values 1-5	CT plausible values 1-5	Mean of scale scores	CIL plausible values 1-5	CT plausible values 1-5
Chile	3092	2.7	4.1	N/A	1128	761	N/A
Denmark	2404	1.6	1.8	1.5	1515	1358	1611
Finland	2546	1.6	2.5	2.2	1589	1037	1157
France	2940	1.6	1.8	1.5	1847	1595	1899
Germany	3655	2.3	3.1	3.0	1578	1186	1232
Italy	2810	1.6	2.3	N/A	1758	1247	N/A
Kazakhstan	3371	3.6	5.4	N/A	934	622	N/A
Korea, Republic of	2875	2.2	2.1	2.9	1325	1368	980
Luxembourg	5401	0.8	0.8	0.6	6445	6779	8447
Portugal	3221	2.1	2.9	2.3	1517	1099	1424
United States	6790	2.3	2.6	2.6	2984	2649	2637
Uruguay	2613	1.7	3.1	N/A	1574	843	N/A
Benchmarking participants							
Moscow (Russian Federation)	2852	1.9	2.2	N/A	1490	1283	N/A
North Rhine-Westphalia (Germany)	1991	1.4	1.8	1.5	1428	1097	1293
ICILS 2018 average	3326	2.0	2.6	2.0	1937	1637	2298

Note: N/A = not available.

Table 6.7: Design effects of main outcome variables—Teacher survey

Country	Sample size	Design effects using scale scores pertaining to students' use of and engagement with ICT at home and school														Mean of scale scores	
		T_USETOOL	T_USEUTIL	T_ICTEFF	T_ICTPRAC	T_CLASACT	T_CODEMP	T_ICTEMP	T_EXLES	T_EXPREP	T_VWNEG	T_VWPOS	T_COLICT	T_PROFSTR	T_PROFREC	T_RESRC	Design effect	Effective sample size
Chile	1686	2.1	1.6	2.0	2.4	2.0	3.0	3.0	2.6	4.0	2.4	3.0	2.1	2.6	1.6	6.8	2.7	617
Denmark	1118	1.5	1.8	1.9	2.5	3.3	1.1	1.5	1.1	1.4	1.2	1.9	2.4	2.6	1.2	6.2	2.1	531
Finland	1853	1.4	1.2	1.3	1.3	2.0	1.1	1.7	1.9	1.3	1.4	1.1	1.3	2.2	1.5	3.3	1.6	1155
France	1462	1.4	1.5	1.4	1.9	1.6	1.2	1.8	2.4	1.6	2.0	1.8	1.4	2.0	1.9	3.6	1.8	804
Germany	2328	1.9	2.1	3.3	2.6	3.2	2.2	3.1	3.4	4.2	2.6	3.3	4.2	4.0	4.2	7.2	3.4	679
Italy	1775	1.8	1.5	1.4	2.2	3.1	1.3	1.6	1.7	2.2	1.7	1.9	2.0	2.2	1.9	3.2	2.0	891
Kazakhstan	2623	9.7	7.1	2.5	6.3	10.5	4.5	5.4	4.3	4.7	5.1	3.9	5.7	8.4	6.4	10.8	6.3	413
Korea, Republic of	2127	3.1	3.1	2.0	1.7	5.2	2.6	2.5	2.3	1.9	2.9	2.0	1.7	3.2	1.4	2.8	2.6	829
Luxembourg	494	1.0	0.9	1.0	1.0	1.2	1.3	1.3	0.7	1.2	1.1	0.8	1.5	1.2	1.6	0.9	1.1	439
Portugal	2823	2.3	2.8	2.7	2.9	2.4	1.2	1.7	3.7	2.3	2.1	2.1	2.2	2.0	2.0	4.5	2.5	1146
United States	3218	4.4	3.6	3.7	2.3	3.5	4.1	2.5	3.5	3.7	2.4	3.1	3.9	3.4	4.7	4.9	3.6	896
Uruguay	1320	1.4	0.9	1.5	1.5	1.7	1.1	1.0	1.5	1.6	1.1	1.4	0.9	1.5	1.7	2.8	1.4	922
Benchmarking participants																		
Moscow (Russian Federation)	2235	4.3	3.6	2.4	1.9	2.4	1.8	1.6	2.1	3.2	2.8	2.1	2.4	3.1	2.2	4.1	2.7	838
North Rhine-Westphalia (Germany)	1468	1.4	2.0	2.1	1.5	1.3	0.8	1.3	2.1	0.9	1.3	1.2	3.1	2.2	1.7	4.6	1.8	798
ICILS 2018 average	1895	2.7	2.4	2.1	2.3	3.1	1.9	2.1	2.4	2.4	2.2	2.1	2.5	2.9	2.4	4.7	2.6	783

Scale scores

T_USETOOL	Use of digital learning tools
T_USEUTIL	Use of general utility software
T_ICTEFF	eachers ICT self-efficacy
T_ICTPRAC	Use of ICT for teaching practices in class
T_CLASACT	Use of ICT for classroom activities
T_CODEMP	Teacher emphasis of teaching coding tasks in class
T_ICTEMP	Emphasis on ICT capabilities in class
T_EXLES	ICT experience with ICT use during lessons
T_EXPREP	ICT experience with ICT use for preparing lessons
T_VWNEG	Negative views on using ICT in teaching and learning
T_VWPOS	Positive views on using ICT in teaching and learning
T_COLICT	Collaboration between teachers in using ICT
T_PROFSTR	Teacher participation in structured learning professional development related to ICT
T_PROFREC	Teacher participation in reciprocal learning professional development related to ICT
T_RESRC	Availability of computer resources at school

References

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*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030387808

Meinck, S. (2015). Sampling design and implementation. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *International Computer and Information Literacy Study 2013 technical report* (pp. 67-86). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Joncas, M., & Foy, P. (2012). Sample design and implementation. In M.O. Martin, & I.V.S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Rust, K., & Adams, R. J. (1999). *Technical standards for IEA studies*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Meinck, S., & Vandenplas, C. (2012). Evaluation of a prerequisite of hierarchical linear modeling (HLM) in educational research - The relationship between the sample sizes at each level of a hierarchical model and the precision of the outcome model. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, Special Issue 1*, October 2012.

UNESCO. (2006). *International Standard Classification of Education: ISCED 1997*. Montreal, Canada: UNESCO Institute for Statistics.

Zuehlke, O. (2011). Sample design and implementation. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS technical report*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

CHAPTER 7:

Sampling weights, non-response adjustments, and participation rates

Sabine Tieck

Introduction

One major objective of ICILS 2018 was to make use of data collected from samples of students' teachers and schools within each country (or educational 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 not provide unbiased estimates of the corresponding population 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 2018, this characteristic applied to the sampled students, teachers, and schools. Furthermore, non-participation has the potential to bias results, 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.[1] All findings presented in ICILS 2018 reports are based on weighted data. Anyone conducting secondary analysis to report on population estimates should account for this in their analyses.

This chapter is largely based on Chapter 7 of the ICILS 2013 technical report (Meinck and Cortes 2015). It describes the conditions under which students, teachers, and schools were considered "participants." Descriptions of how the several sets of weights and non-response adjustments were computed follow. Please note that Chapter 13 of this report covers the use of the jackknife replication method (needed for variance estimation). Subsequent sections describe the computation of participation rates at each sampling stage, and the minimum participation requirements. The ICILS 2018 research team regarded response rates as an important indicator of data quality and the achieved quality of sample implementation for each country is presented.

Types of sampling weights

The ICILS 2018 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 non-participation 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 schools was selected. If some of the sampled schools refused to participate, then the remaining (participating) schools in this stratum would carry the (base) weight of the non-participating

1 For further reading on the topic, we recommend Meinck (2015), Rust (2014), Groves et al. (2004), and Statistics Canada (2003).

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 2018 within a stratum).

Calculating student 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. When explicit stratification was used, the school samples were selected independently from within each explicit stratum h, with h =1,..., 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 small (M_{hi}<20), the measure of size M_{hi} was redefined as the average size of all small schools in that stratum. In a few countries, equiprobable systematic random sampling (SyRS) was applied in particular strata.

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:

$$WGTFAC1_{hi} = \frac{M_h}{n_h^s \times M_{hi}} \quad \text{for PPS sampling and}$$

$$WGTFAC1_{hi} = \frac{N_h}{n_h^s} \quad \text{for SyRS}$$

where n_h^s is the number of sampled schools in stratum h, M_h is the total number of students enrolled in the schools of explicit stratum h, M_{hi} is the measure of size of the selected school i, and N_h is the total number of schools in stratum h.

School non-response adjustment (WGTADJ1S)

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-school 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:

$$WGTADJ1S_{hi} = \frac{n_h^{s,e}}{n_h^{p\text{-}std}}$$

where $n_h^{s,e}$ is the number of sampled eligible schools and $n_h^{p\text{-}std}$ is the number of participating schools (whether originally sampled or replacement schools) in the student survey 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 2018. 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, or had only excluded students enroled. Ineligible schools such as these were not taken into account when calculating the non-response adjustment.

Student base weight (WGTFAC3S)

The IEA Windows Within-School Sampling Software (IEA WinW3S) was used to manage within-school sampling during the second sampling stage (see Chapter 8 for details) in order to conduct a systematic random selection of students from the target grade. The student base weight for student *k* was calculated as:

$$WGTFAC3S_{hik} = \frac{M_{hi}}{m_{hi}}$$

where M_{hi} is the total number of students in the target grade in school *i* in stratum *h* and m_{hi} is the number of sampled students in school *i* in stratum *h*.

In schools with fewer than 26 target grade students, all eligible students were selected for participation. In these cases, the weight factor was set to a value of one.[2]

Student non-response adjustment (WGTADJ3S)

Unfortunately, not all selected students were able or willing to participate in ICILS 2018. To account for the reduction in sample size due to within-school non-participation, a student non-response adjustment factor was introduced. 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 within a surveyed school. Accordingly their sampling weights had to be adjusted.

The adjustment for student non-response for each participating student *k* was calculated as:

$$WGTADJ3S_{hik} = \frac{m^{e}_{hi}}{m^{p}_{hi}}$$

with m^{e}_{hi} being the number of eligible students in school *i* in stratum *h* and m^{p}_{hi} being the number of participating students in school *i* in stratum *h*. In the context of student weight adjustment, students of the target population were regarded as eligible if they had not been excluded due to disabilities or language problems.[3]

Please note that sampled students who did not participate in the survey because they had left the sampled school after within-school 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 school *i* in stratum *h* is the product of the four student-weight components:

$$TOTWGTS_{hik} = WGTFAC1_{hi} \times WGTADJ1_{hi} \times WGTFAC3_{hik} \times WGTADJ3_{hik}$$

2 Two countries deviated from this rule: in Luxembourg, all students were sampled, and in the Unites States, the sample size for the students was set to 30.

3 For Chile this adjustment factor includes a gender adjustment factor as the original population estimates regarding the distribution of girls and boys did not match the recorded proportion of boys and girls in Chile (although the estimated total number of students did match the recorded population figures). This could be because the distribution of single-sex schools was not controlled for in the sample. The adapted adjustment factor fits the population estimates regarding the population size and the gender distribution.

Calculating teacher weights

School base weight (WGTFAC1)

As 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 study 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*:

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

Here, $n_h^{s,e}$ 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:

$$WGTFAC2T_{hil} = \frac{T_{hi}}{t_{hi}^s}$$

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 fewer than 21 target grade teachers, all eligible teachers were selected for participation. In these cases this weight factor was equal to one.[4]

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

$$WGTADJ2T_{hil} = \frac{t_{hi}^{s,e}}{t_{hi}^p}$$

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.

4 In Luxembourg, the number of teachers to select was increased to 20, thus all eligible teachers were selected in schools with less than 25 target grade teachers.

Teacher multiplicity factor (WGTFAC3T)

Some teachers in ICILS 2018 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:

$$WGTFAC3T_{hil} = \frac{1}{f_{hil}}$$

Here, 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:

$$TOTWGTT_{hil} = WGTFAC1_{hi} \times WGTADJ1T_{hi} \times WGTFAC2T_{hil} \times WGTADJ2T_{hil} \times WGTFAC3T_{hil}$$

Calculating school weights

ICILS 2018 was designed as a survey of students and teachers but not as a school survey. However, in order to collect background information at school level, a principal questionnaire and an ICT coordinator questionnaire were administered to every participating school. School weights were calculated and included in the international database in order to allow for analyses at the school level. However, results at the school level should be interpreted with some caution as they may be subject to considerable sampling error.

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 *i* as follows:

$$WGTADJ1C_{hi} = \frac{n_h}{n_h^{p-sch}}$$

Here, 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 this kind of 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:

$$TOTWGTC_{hil} = WGTFAC1_{hi} \times WGTADJ1C_{hi}$$

Calculating participation rates

For ICILS 2018, 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 is calculated as:

$$UPRS_{schools\_BR} = \frac{n_{op}}{n_{fp} + n_{np}}$$

The unweighted school participation rate in the student survey *after* replacement is computed as:

$$UPRS_{schools\_AR} = \frac{n_{fp}}{n_{fp} + n_{np}}$$

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:

$$UPRS_{overall\_BR} = UPRS_{schools\_BR} \times UPRS_{students}$$

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

$$UPRS_{overall\_AR} = UPRS_{schools\_AR} \times UPRS_{students}$$

Weighted participation rates in the student survey

The weighted school participation rate in the student survey *before* replacement was calculated as the ratio of summations of all participating students *k* in stratum *h* and school *i*:

$$WPRS_{schools\_BR} = \frac{\sum_h \sum_{i \in op} \sum_{k \in sfp} WGTFAC1_{hi} \times WGTFAC3S_{hik} \times WGTADJ3S_{hik}}{\sum_h \sum_{i \in fp} \sum_{k \in sfp} WGTFAC1_{hi} \times WGTADJ1S_{hi} \times WGTFAC3S_{hik} \times WGTADJ3S_{hik}}$$

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 is therefore:

$$WPRS_{schools\_AR} = \frac{\sum_h \sum_{i \in fp} \sum_{k \in sfp} WGTFAC1_{hi} \times WGTFAC3S_{hik} \times WGTADJ3S_{hik}}{\sum_h \sum_{i \in fp} \sum_{k \in sfp} WGTFAC1_{hi} \times WGTADJ1S_{hi} \times WGTFAC3S_{hik} \times WGTADJ3S_{hik}}$$

The weighted student participation rate was computed as follows, taking again replacement schools into account:

$$WPRS_{students} = \frac{\sum_h \sum_{i \in fp} \sum_{k \in sfp} WGTFAC1_{hi} \times WGTFAC3S_{hik}}{\sum_h \sum_{i \in fp} \sum_{k \in sfp} WGTFAC1_{hi} \times WGTFAC3S_{hik} \times WGTADJ3S_{hik}}$$

The weighted overall participation rate in the student survey *before* replacement is therefore:

$$WPRS_{overall\_BR} = WPRS_{schools\_BR} \times WPRS_{students}$$

The weighted overall participation rate in the student survey *after* replacement is:

$$WPRS_{overall\_AR} = WPRS_{schools\_AR} \times WPRS_{students}$$

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. Differences between the two tables indicate different response patterns among strata with disproportional sample allocations. For example, the unweighted school participation rate of Germany was considerably higher than the weighted rate because the federal state of North Rhine-Westphalia, in which almost all schools participated, was oversampled for Germany. North Rhine-Westphalia participated in ICILS both as a state within Germany and as a benchmarking participant. In comparison, relatively fewer schools participated in the remaining strata across Germany.

It should be noted that only those schools that had at least a participation rate of 50 percent among their sampled students were treated as participants in the student survey. 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 and student participation rates—Student survey

Country	School participation rate (%)		Student participation rate (%)	Overall participation rate (%)	
	Before replacement	After replacement		Before replacement	After replacement
Chile	91.1	99.4	93.6	85.2	93.1
Denmark	76.0	95.3	85.3	64.8	81.3
Finland	97.9	98.6	91.8	89.9	90.6
France	99.4	100.0	94.7	94.1	94.7
Germany	84.3	91.3	88.5	74.6	80.8
Italy	95.3	100.0	95.0	90.6	95.0
Kazakhstan	99.5	99.5	97.9	97.3	97.3
Korea, Republic of	100.0	100.0	96.7	96.7	96.7
Luxembourg	92.7	92.7	90.1	83.5	83.5
Portugal	86.3	91.3	81.1	70.0	74.1
United States	67.5	76.9	90.8	61.4	69.9
Uruguay	91.9	95.9	80.6	74.0	77.3
Benchmarking participants					
Moscow (Russian Federation)	98.7	100.0	95.8	94.5	95.8
North Rhine-Westphalia (Germany)	92.9	97.3	91.6	85.0	89.1

Table 7.2: Weighted school and student participation rates—Student survey

Country	School participation rate (%)		Student participation rate (%)	Overall participation rate (%)	
	Before replacement	After replacement		Before replacement	After replacement
Chile	91.0	100.0	93.1	84.8	93.1
Denmark	75.6	95.3	84.8	64.1	80.8
Finland	98.3	98.6	91.9	90.3	90.6
France	99.4	100.0	95.0	94.4	95.0
Germany	78.9	88.3	86.6	68.3	76.5
Italy	95.1	100.0	94.9	90.3	94.9
Kazakhstan	99.5	99.5	97.6	97.2	97.2
Korea, Republic of	100.0	100.0	96.7	96.7	96.7
Luxembourg	96.4	96.4	90.1	86.9	86.9
Portugal	85.7	90.2	80.0	68.6	72.2
United States	67.4	77.1	91.0	61.4	70.2
Uruguay	90.7	95.7	80.2	72.8	76.8
Benchmarking participants					
Moscow (Russian Federation)	98.2	100.0	95.7	93.9	95.7
North Rhine-Westphalia (Germany)	92.6	97.4	91.0	84.2	88.6

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:

$$WPRT_{schools\_BR} = \frac{\sum_h \sum_{i \in op} \sum_{l \in tfp} WGTFAC1_{hi} \times WGTFAC2T_{hil} \times WGTADJ2T_{hil} \times WGTFAC3T_{hil}}{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGTFAC1_{hi} \times WGTADJ1T_{hi} \times WGTFAC2T_{hil} \times WGTADJ2T_{hil} \times WGTFAC3T_{hil}}$$

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

$$WPRT_{schools\_AR} = \frac{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGTFAC1_{hi} \times WGTFAC2T_{hil} \times WGTADJ2T_{hil} \times WGTFAC3T_{hil}}{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGTFAC1_{hi} \times WGTADJ1T_{hi} \times WGTFAC2T_{hil} \times WGTADJ2T_{hil} \times WGTFAC3T_{hil}}$$

The weighted teacher participation rate is therefore:

$$WPRT_{teachers} = \frac{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGTFAC1_{hi} \times WGTFAC2T_{hil} \times WGTFAC3T_{hil}}{\sum_h \sum_{i \in fp} \sum_{l \in tfp} WGTFAC1_{hi} \times WGTFAC2T_{hil} \times WGTADJ2T_{hil} \times WGTFAC33T_{hil}}$$

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. As described earlier, Germany provides a prominent example of this effect.

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 rates—Teacher survey

Country	School participation rate (%)		Teacher participation rate (%)	Overall participation rate (%)	
	Before replacement	After replacement		Before replacement	After replacement
Chile	89.9	97.2	93.5	84.1	90.9
Denmark	72.7	92.0	84.4	61.4	77.7
Finland	97.3	97.9	92.2	89.7	90.3
France	78.2	78.2	81.0	63.4	63.4
Germany	73.8	79.5	85.3	62.9	67.8
Italy	94.0	98.7	92.8	87.3	91.6
Kazakhstan	100.0	100.0	99.9	99.9	99.9
Korea, Republic of	100.0	100.0	100.0	100.0	100.0
Luxembourg	68.3	68.3	75.0	51.2	51.2
Portugal	90.9	95.9	91.4	83.0	87.6
United States	66.1	75.7	88.6	58.5	67.1
Uruguay	66.9	70.3	75.3	50.4	53.0
Benchmarking participants					
Moscow (Russian Federation)	98.7	100.0	100.0	98.7	100.0
North Rhine-Westphalia (Germany)	91.1	95.5	90.3	82.3	86.3

Table 7.4: Weighted school and teacher participation rates—Teacher survey

Country	School participation rate (%)		Teacher participation rate (%)	Overall participation rate (%)	
	Before replacement	After replacement		Before replacement	After replacement
Chile	91.2	96.9	93.6	85.3	90.7
Denmark	70.4	92.0	84.0	59.2	77.3
Finland	97.8	98.0	92.5	90.4	90.7
France	78.4	78.4	80.6	63.2	63.2
Germany	63.1	70.5	81.7	51.5	57.5
Italy	93.8	98.6	91.9	86.2	90.6
Kazakhstan	100.0	100.0	100.0	100.0	100.0
Korea, Republic of	100.0	100.0	100.0	100.0	100.0
Luxembourg	68.5	68.5	75.6	51.8	51.8
Portugal	89.0	95.3	91.6	81.5	87.3
United States	62.2	72.4	89.4	55.6	64.7
Uruguay	69.5	74.1	74.5	51.8	55.2
Benchmarking participants					
Moscow (Russian Federation)	97.6	100.0	100.0	97.6	100.0
North Rhine-Westphalia (Germany)	90.2	95.6	91.1	82.2	87.2

ICILS 2018 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 2018 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 2018 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 biasing results. Because very little information about non-respondents was 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 2018 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. 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 overall amount of partial non-response (i.e., omitted items in questionnaires or tasks that had not been attempted) was minimal.

There is evidence that attendance and academic performance tend to be positively correlated (Balfanz and 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 2018 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 low.

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.

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.

Country-level participation requirements

Three categories for sampling participation were defined:

- Countries grouped in Category 1 met the ICILS 2018 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 2018 sampling participation requirements.

Sampling participation categories for the teacher survey were identical to the ones in the student survey. The results from ICILS 2018 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 will be 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 had 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 will be placed in Category 3.

Reporting data

The ICILS 2018 research team considered it necessary to make readers of the international report 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.

For the student survey, nine countries and both benchmarking participants were reported as meeting sampling participation requirements in Category 1 and were reported without annotation, while six countries and the two benchmarking participants were in this category for the teacher survey. Two countries were in Category 2 for the student survey[5] and one country was in this category for the teacher survey. One country had its student survey results reported in a separate section of the tables as a Category 3 country, while for the teacher survey this was the case for five countries. All ICILS 2018 countries and benchmarking participants had participation rates of their originally sampled schools above 50 percent.

Table 7.5 lists the participation categories of each country for the student and the teacher surveys.

Table 7.5: Achieved participation categories by country

Country	Participation category	
	Student survey	Teacher survey
Chile	1	1
Denmark	2	2
Finland	1	1
France	1	3
Germany	1	3
Italy	1	1
Kazakhstan	1	1
Korea, Republic of	1	1
Luxembourg	1	3
Portugal	2	1
United States	3	3
Uruguay	2	3
Benchmarking participants		
Moscow (Russian Federation)	1	1
North Rhine-Westphalia (Germany)	1	1

5 Please note that Portugal is reported for the student survey in category two, because they only slightly missed the required minimum parcipation rate.

References

Balfanz, R., & Byrnes, V. (2012). *Chronic absenteeism: Summarizing what we know from nationally available data*. Baltimore, MD: Johns Hopkins University Center for Social Organization of Schools.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York, NY: Wiley.

Hancock, K. J., Shepherd, C. C. J., Lawrence, D., & Zubrick, S. R. (2013). *Student attendance and educational outcomes: Every day counts*. Canberra, Australia: Report for the Department of Education, Employment and Workplace Relations.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Meinck, S. (2015). Computing sampling weights in large-scale assessments in education. *Survey methods: Insights from the field*. http://surveyinsights.org/?p=5353

Meinck, S., & Cortes, D. (2015). Sampling weights, nonresponse adjustments, and participation rates. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt, *International Computer and Literacy Information Study 2013 technical report* (pp. 87-112). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Statistics Canada. (2003). *Survey methods and practices*. Ottawa, Canada: Author. https://www150.statcan.gc.ca/n1/pub/12-587-x/12-587-x2003001-eng.pdf

Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In Rutkowski, L., von Davier, M., & Rutkowski, D., (Eds), *Handbook of international large-scale assessment*. New York: CRC Press.

CHAPTER 8:

ICILS 2018 field operations

Ekaterina Mikheeva and Sebastian Meyer

Introduction

Successful administration of ICILS 2018 assessment depended heavily on the contributions of the study's national research coordinators (NRCs) and national center staff. As is the situation for all large-scale cross-national surveys, administration of the assessment along with the overall coordination and logistical aspects of the study presented a set of significant challenges for each participating country. These challenges were heightened by the demands of administering the ICILS 2018 student instruments on computer.

The ICILS international study center (ISC) at ACER in cooperation with IEA therefore developed internationally standardized field operations procedures to assist the NRCs and to aid uniformity of their instrument-administration activities. The international team designed these procedures to be flexible enough to simultaneously meet the needs of individual participants and the high quality expectations of IEA survey standards. The team began by referring to the ICILS 2013 study procedures and those used in other IEA studies, such as IEA's Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS), and International Civic and Citizenship Education Study (ICCS), and then tailoring these to suit the specific requirements of ICILS 2018, most importantly the computer-based administration of the student instruments.

All national centers received guidelines on the survey operations procedures for each stage of the assessment. The guidelines advised on contacting schools, listing and sampling students, preparing materials for data collection, administering the assessment, scoring the assessment, and creating data files. National centers also received materials on procedures for quality control and were asked to complete online questionnaires that asked for feedback on the survey activities.

Field operations personnel

The role of the national research coordinators and their centers

One of the first steps that all countries or education systems participating in ICILS 2018 had to take when establishing the study in their country was to appoint an NRC. The NRC acted as the main contact person for all those involved in ICILS 2018 within the country and was the country representative at the international level.

NRCs 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 international project staff and national experts.

The role of school coordinators and test administrators

In order to facilitate successful administration of ICILS 2018, the international team required the establishment of two roles 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.

In cooperation with school principals, national centers identified and trained school coordinators for all participating schools. The school coordinator could be a teacher or other staff member in the school. The school coordinator could also be the test administrator at the school, but was not to be a teacher of any of the sampled students. In some cases, national centers appointed external individuals as school coordinators. The coordinators' responsibilities included the following major tasks:

- Identifying eligible students and 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 instruments and 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 testing; and
- Ensuring that the test administrators return all testing materials after the testing session.

The test administrators were mainly responsible for administering the student test and questionnaire. They were employed either by the national center or directly by the schools. Accordingly, 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.

Field operations resources

Manuals and documentation

The international study team released the ICILS 2018 survey operations procedures manuals to the NRCs in five units, each of which was accompanied by additional materials, including manuals for use in schools and software packages. All of this material was organized and distributed chronologically according to the stages of the study.

The five units and their accompanying manuals and software packages were:

- *Unit 1: Sampling Schools* specified the actions and procedures required to develop a national sampling plan in compliance with the international ICILS 2018 sample design.
- *Unit 2: Working with Schools* contained information about how to work with schools in order to plan for successful administration of the ICILS 2018 instruments.
- *Unit 3: Instrument Preparation* described the processes involved in preparing the ICILS 2018 instruments for production and use in countries.
- *Unit 4: Data Collection and Quality Monitoring Procedures* dealt with the processes involved in preparing for, supporting, and monitoring ICILS 2018 data collection in schools.
- *Unit 5: Post Collection Data Capture, Data Upload, Scoring and Parental Occupation Coding* provided guidelines on post-data collection processes and tasks. These included, but were not limited to, data capture from the paper questionnaires, uploading student assessment data, scoring student responses, and coding parental occupations.
- The *School Coordinator Manual*, subject to translation, described the role and responsibilities of the school coordinator, the main contact person within each participating school.
- The *Test Administrator Manual*, subject to translation, described the role and responsibilities of the test administrator, whose work included administration of the student assessment.
- The *National Quality Observer Manual* provided national quality control observers with information about ICILS 2018, their role and responsibilities during the project, and the timelines, actions, and procedures to be followed in order to carry out the national quality control programs.

- The *International Quality Observer Manual* provided international quality observers with information about ICILS 2018, their role and responsibilities during the project, and the timelines, actions, and procedures to be followed in order to carry out the international quality observer programs.
- The *Scoring Guides for Constructed-Response Items*, subject to translation, provided detailed and explicit guidelines on how to score each constructed-response item.
- The *Compatibility Check and School Computer Resources Survey: Instructions for NRCs and Preparing Computers for ICILS – Instructions for School Coordinators Manual* addressed whether computers in the sampled schools could be used for the ICILS 2018 assessment and whether special arrangements needed to be made in order to administer the assessment.

Software

The international project team also supplied NRCs with software packages to assist with data collection and scoring of constructed-response items:

- *IEA Translation System*: This web-based application supported translation, translation verification, and layout verification of the teacher, principal, and ICT coordinator questionnaires. This software also allowed for cultural adaptations and verification for English-speaking countries that did not require translations.
- *AM Designer (RM Results)*: This web-based application supported translation, translation verification, and layout verification of the student instruments (test modules, tutorial, and questionnaire). This software also allowed for cultural adaptations and verification for English-speaking countries that did not require translations.
- *AM Examiner*: This software was used to administer the computer-based ICILS 2018 test and contextual questionnaire to the students. The software was run from USB sticks either on existing computers in the school or on a set of laptop computers provided specifically for the ICILS 2018 administration. Alternatively, a laptop server administration method was used when it was not possible to run the test software on USB sticks connected to individual computers.
- *AM Marker*: This web-based application enabled scoring administrators, scoring team leaders, and scorers to manage and carry out the scoring process for constructed-response items. The software allowed NRCs to create scoring teams and to assign scorers to scoring teams. The software also included a training tool that enabled navigation between sections, score training responses, score student responses, and flag responses for review scoring.
- *IEA Coding Expert*: This client-based application enabled coding administrators, coding team leaders, and coding staff to manage and carry out the coding process for open-ended student responses with respect to parents' occupation in the student questionnaire. The software allowed NRCs to create teams and to assign staff to coding teams. The software also included a training tool that enabled coding of student responses and flagging responses for review coding.
- *IEA Windows Within-School Sampling Software* (IEA WinW3S): This enabled the ICILS 2018 national centers to select students 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* (IEA OSS): This software enabled verified text passages in the questionnaires to be transferred from the IEA Translation System to online questionnaires, with these online versions then delivered to respondents.

- *IEA Data Management Expert* (IEA DME): This software facilitated the entering of paper questionnaire data. The IEA DME software also allowed national adaptations to be made to the questionnaires and provided a set of data quality control checks.

In addition to preparing the software and manuals, IEA conducted data-management training designed to train national center staff on all procedures and the software supporting these procedures. Namely, IEA WinW3S, IEA and RM Results translation systems, IEA DME, and the AM Examiner. This seminar was combined with a scoring training, during which national center staff were trained to use the AM Marker. Instructions for using the ICILS translation systems were covered in one of the regular NRC meetings.

Field operations processes

Linking students and teachers to schools

The international project staff established a system to assign hierarchical identification codes (IDs). These uniquely identified and allowed tracking of the sampled schools, teachers, and students. Table 8.1 represents the hierarchical identification system codes.

Every sampled student was assigned an eight-digit ID number unique within each country. Each number consisted of the four-digit number identifying the school, followed by a two-digit number identifying the student group within the school (01 for all), and a two-digit number identifying the student within that group.

Each sampled target-grade teacher was assigned a teacherID number consisting of the four-digit school number followed by a two-digit teacher number 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
Student	School (C), Student Group (G, constant: 01), Student (S)	CCCCGGSS	10010101
Teacher	School (C), Teacher (T)	CCCCTT	100101

Activities for working with schools

In ICILS 2018, the within-school sampling process and the assessment administration required close cooperation between the national centers and representatives from the schools, that is, 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 NRCs had obtained a list of the schools sampled for ICILS 2018 (for more information on sampling procedures, please refer to Chapter 6 of this report), it was important for the success of the study that national centers established good working relationships with the selected schools. NRCs were responsible for contacting the schools and encouraging them to take part in the assessment, a process that often involved obtaining support from national or regional educational authorities or other stakeholders, depending on the national context.

In cooperation with school principals, national centers identified and trained school coordinators for all participating schools. The school coordinator could be a teacher or guidance counselor in the school. In cases where the school coordinator also acted as the test administrator at the school, he or she was not allowed to be a teacher of the sampled class. In some cases, national centers appointed one of their own members to fill this role. Often this person was responsible for several schools in an area. Each school coordinator was provided with an *ICILS School Coordinator Manual*, which described their responsibilities in detail and encouraged them to contact the national center if they had any questions.

School coordinators were required to provide all required information about their respective schools and additionally coordinate 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 the USB-based *compatibility check*, and send results to the national study center. School coordinators were also responsible for obtaining parental permission as necessary, liaising with the test administrator to coordinate the test session, distributing 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.

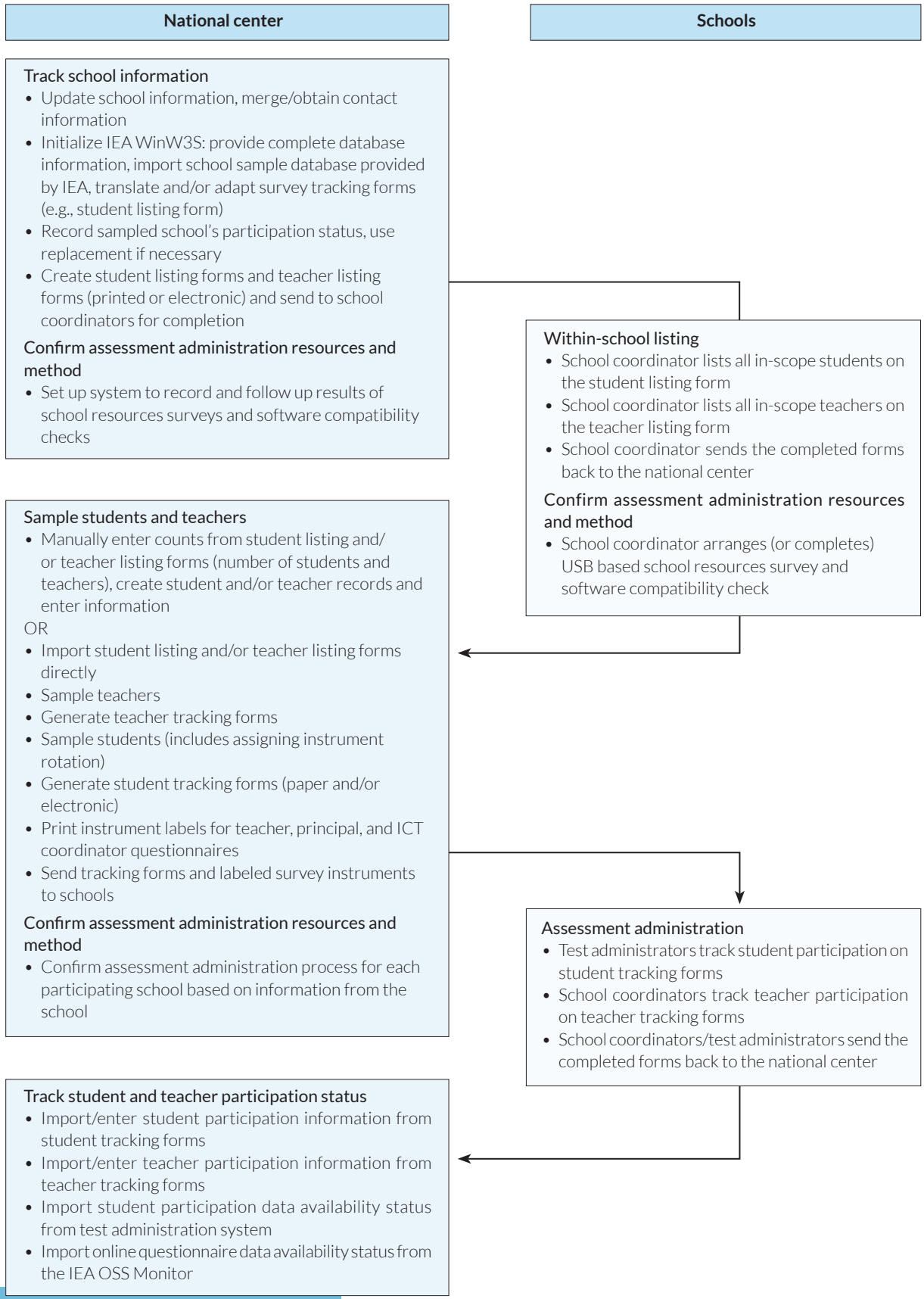
National centers sent a *student listing form* to each school coordinator and asked them to provide information on all the eligible target-grade students in the school. School coordinators collected details about these students, such as their names (if country regulations allowed names to be provided to the national centers), birth month and year, gender, exclusion status,[1] and the assessment language of the student (in case the national center provided different language versions of the student instruments).

The national centers used this information to sample students within the schools. Listing all eligible students in the target grade was key to ensuring that every student in the target population had a known chance of being sampled, an essential requirement for obtaining random samples from all of the target-grade students at and across schools.

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.

1 Although all students enrolled in the target grade were part of the target population, ICILS 2018 recognized that some student exclusions were necessary because of a physical or intellectual disability, or in cases of non-native language speakers without the language proficiency to complete the assessment. Accordingly, the sampling guidelines allowed for the exclusion of students with any of several disabilities (for more information on sampling procedures, please see Chapter 6). Countries were required to track and account for all students, yet flagged those for which exemptions were defined. Because the local definition of such disabilities could vary from country to country, it was important that the conditions under which countries excluded students were carefully documented.

Figure 8.1: Activities with schools



Preparing the computer-based test delivery at schools

Because ICILS 2018 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.

The compatibility check and school computer resources survey were administered in order to answer two 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; and (ii) if, in those cases where the school computers could be used for testing, special arrangements would be needed (e.g., altering the configuration of computers or using a laptop with the local server connected to the school LAN) for the USB-based student test to run correctly.

The process of administering the compatibility check and school resources survey required NRCs in non-English speaking countries to translate the school computer resources survey questions and to make them available along with the USB compatibility check file to school coordinators on a USB stick.

After receiving the USB sticks containing the compatibility check files and instructions, school coordinators were required to:

- Run the USB compatibility check on every computer that was to be used for the ICILS 2018 assessment;
- Complete one of the included school computer resources surveys per school; and
- Send the results back to the national study center.

This information on the availability and compatibility of the participating schools' computers enabled national centers to determine the best test delivery method for each school.

The national centers then sent the following items to each school: the necessary tracking forms, labels, questionnaires (online or paper-based), and manuals as well as USB sticks matching the number of students listed on the student tracking form (plus three extra sticks).

Administering the assessment at schools

The process of distributing the printed materials and the electronic student instruments to the schools required the national centers to engage in careful organization and planning.

The national centers sent teacher questionnaires to each teacher listed on the teacher tracking form, in each school. They also sent a principal questionnaire to each school's principal and an ICT coordinator questionnaire to each school's ICT coordinator.

The national centers furthermore prepared and sent cover letters containing login information and instructions on how to complete the online questionnaire to all teachers, school principals, and ICT coordinators who had elected to complete their questionnaires online. National center staff sent the packaged materials to the school coordinators prior to the testing date and asked them to confirm the receipt of all instruments. School coordinators then distributed the school questionnaire and teacher questionnaires (or the cover letters for the online participants) while ensuring that the other instruments were kept in a secure room until the assessment date.

In accordance with the international guidelines and requirements as well as local conditions, national centers assigned a test administrator to each school. In some cases, the school coordinator also acted as the test administrator. The test administrators received training from the national centers. 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 server method); leading students through the assessment; and, accurately timing the sessions.

The student tracking forms indicated, for each sampled student, the assigned student instrument, which consisted of the two test-item modules and the student questionnaire, administered via the ICILS 2018 Student Test Software. For ICILS 2018 countries administering computer and information literacy (CIL) test modules only (i.e., Chile, Italy, Kazakhstan, Uruguay, and Moscow, Russian Federation), administration of the assessment consisted of three parts, the first two of which required students to complete CIL test modules and the third to answer the student questionnaire. In countries participating in the computational thinking (CT) test modules (i.e., Denmark, Finland, France, Germany, Korea, Luxembourg, Portugal, the United States, and North Rhine-Westphalia, Germany), the test session consisted of five parts with both CT test modules following the two CIL test modules and student questionnaire parts. Test administrators were requested to document student participation on the student tracking forms.

During the administration of the assessment test, administrators were required to provide a range of instructions to students. When administering 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.

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 (as described above), the test session was extended by an additional 25 minutes per 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.

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

As in the previous cycle, ICILS 2018 offered participating countries the option of administering the principal, ICT coordinator, and teacher questionnaires online instead of in paper form. To ensure comparability of the data from the online and the paper modes, only those countries that had previously tested the online data collection during the ICILS 2018 field trial were allowed to use the online option during the main survey. All countries used the online administration mode for their schools.

After the principal, ICT coordinator, and teacher questionnaires had gone through the translation and translation verification processes, they were prepared for delivery online using the IEA Online Survey System (IEA OSS) software as described in more detail in Chapter 5 of this report.

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.)**

The IEA OSS is a hierarchical model of a survey that stores and manages all questionnaire-related information, including text passages, translations and adaptations, verification rules, variable names, and information for data management.

To serve the different possible usage scenarios, the IEA OSS comprises three distinct components. The *Designer* component was used to create, delete, disable, and edit survey components (e.g., questions and categories) and their properties. It enabled translation of all text passages in the existing national paper questionnaires and additional system texts, and it included a complete web server to verify and preview the survey exactly as if under live conditions. The Designer also supported the export of codebooks to IEA's generic data entry software, the IEA DME, to enable isomorphic data entry of online and paper questionnaires. The *Web* component was a compiled application that provided questionnaires in HTML format to the respondents for completion within standard internet browsers. Finally, the web-based *Monitor* component allowed national centers to audit participation in real-time. It also allowed the centers to follow up with schools when questionnaires were incomplete or not returned in a similar way to that used in the administration of the paper questionnaires. The live systems were hosted on dedicated high-performance servers rented from a reliable and experienced solution provider in Germany.

The electronic versions of the ICILS 2018 principal, ICT coordinator, and teacher questionnaires could only be completed via the internet. Accordingly, the design ensured that online 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.

To limit the administrative burden and necessary communication with schools, national centers made the initial decision on whether to assign the online or paper questionnaire as a default to respondents. This decision was based on the centers' and schools' prior experience of participation in similar surveys and during the ICILS field trial. Usually, every respondent in a particular school was assigned the same mode, either online or paper. However, national centers were requested to take into account the mode that a specific school or a particular individual preferred. National

centers had to ensure that every respondent assigned to the online mode by default had the option to request and complete a paper questionnaire, regardless of the reasons for being unwilling or unable to answer online.

To ensure confidentiality and separation, every respondent received individual login information. The national centers sent this information, along with general information on how to access the online questionnaire, to respondents in the form of "cover letters." In line with the procedures used during distribution of the paper questionnaires, the school coordinator delivered this information to the designated individuals.

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 IEA if unable to solve a problem locally.

The navigational structure of the online questionnaire had to be as similar as possible to that of the paper questionnaires. Respondents could use "next" and "previous" buttons to navigate to an adjacent page, as if they were flipping physical pages. In addition, a hypertext "table of contents" mirrored the experience of opening a specific page or question of a paper questionnaire. While most respondents followed the sequence of questions directly, these two features allowed respondents to skip or omit questions, just as if they were answering a self-administered paper questionnaire.

To further ensure the similarity of the two sets of questionnaires, 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.

Certain differences in the representation of the two modes remained, however. To reduce response burden and complexity, the online survey automatically skipped questions not applicable to the respondent, in contrast to the paper questionnaire, which instructed respondents to proceed to the next applicable question. 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.

Although countries using the online mode in ICILS 2018 faced parallel workload and complexity before and during the data collection, they had the benefit of a reduction in workload afterwards. Because answers to online questionnaires were already in electronic format and stored on servers maintained by IEA, there was no need for separate data entry.

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. As the survey activities questionnaire, unlike the other ICILS 2018 questionnaires, did not require national adaptations and was completed in English, it was well suited for online data collection.

The purpose of the survey activities questionnaire was to gather opinions and information about the strengths and weaknesses of the ICILS 2018 assessment materials (e.g., test instruments, manuals, scoring guides, and software) as well as countries' experiences with the ICILS 2018 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.

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.

Scoring the assessment and checking scorer reliability

Scoring the assessment

The success of assessments containing constructed-response items depends on the degree to which student responses are scored reliably. Seventeen of the ICILS 2018 CIL assessment items were constructed-response items, and five large tasks were scored against a total of 37 criteria. One of the large-task criteria was automatically scored by the ICILS scoring system. Human scorers reviewed the automatically generated suggested score and could either accept or modify the score. Of the 102 ICILS 2018 CIL items, 53 were scored by human scorers, and it was critical to the quality of the ICILS 2018 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. There were 17 CT items for ICILS 2018. Two of these items were constructed response and scored by human scorers.

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 2018 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 scoring guides for the three ICILS trend modules were not revised and were identical to those used in the first ICILS cycle. Further details of the development and revision of the ICILS 2018 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 2018 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 scoring system 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 2018 scored data. Scoring reliability within each country required two different scorers to independently score a random sample of 20 percent of 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 20 percent of all responses was 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.

Field trial procedures

The ICILS 2018 field trial was a smaller administration of the ICILS 2018 assessment; on average, approximately 1000 students were tested in each participating country.

The international field trial was conducted from May to June 2017.

The field trial was crucial to the development of the ICILS 2018 assessment instruments and also served the purpose of testing the ICILS 2018 survey operations procedures in order to avoid any possible problems during the ICILS 2018 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. The field trial resulted in some important modifications to survey operations procedures and contributed significantly to the successful implementation of ICILS 2018.

Summary

Considerable efforts were made to ensure high standards of quality in the survey procedures for the ICILS 2018 data collection. NRCs played a key role in implementing the data collection in each participating country, during which they followed internationally agreed upon survey operations procedures. The international study consortium provided NRCs with a comprehensive set of manuals containing detailed guidelines for the preparation of the study, its administration, scoring of open-ended questions, and data processing. National centers also received tailored software packages for sampling and tracking student and teachers within schools, the computer-based student assessment, data capture, and the online administration of contextual questionnaires. The international ICILS 2018 field trial in 2016-2017 was crucial for testing survey operations procedures in participating countries and contributed to the successful implementation of the main data collection.

CHAPTER 9:

Quality assurance procedures for ICILS 2018

Sandra Dohr, Lauren Musu, and David Ebbs

Introduction

Considerable effort was made to develop and standardize materials and procedures for ICILS 2018 in order to collect high-quality and comparable data across countries to the greatest extent possible. For this purpose, quality assurance was an integral part of ICILS 2018 and encompassed all main activities from the ICILS framework. The activities included assessment and questionnaire development, sampling, instrument preparation including the verification of national versions, data collection, scaling, and data analysis. Moreover, quality assurance activities occurred at both international and national levels. Therefore, various members of the consortium and the national study centers were part of the quality assurance components for ICILS 2018. This approach provided the consortium with diverse perspectives, which helped to detect potential issues concerning the survey procedures. It also supported the data adjudication process (see Chapter 11) and provided possible explanations for potential inconsistencies or irregularities in the data.

The aim of this chapter is to provide an overview and present results of quality assurance activities that occurred during the data collection for ICILS 2018 and survey activities on a national level. Precisely, this chapter focuses on the two following aspects of the overall quality assurance activities in ICILS 2018:

- *International quality control program:* The consortium developed and implemented an international quality control program to monitor data collection activities and compliance of national study centers with the procedural standards. For this purpose, IEA contracted and trained independent experts for each participating country, named international quality observers (IQOs). Each IQO visited 15 schools selected from the school sample in the respective country to observe an ICILS testing session, interview the school coordinator and test administrator, and document their observations.
- *Survey activities questionnaire:* The consortium developed a comprehensive questionnaire to gain insights into the national activities before, during, and after the administration of ICILS 2018. National research coordinators (NRCs) had to complete the questionnaire, which covered 13 different areas, including sampling, communication with schools, instrument preparation, data collection, scoring, and data submission. NRCs also reported on national quality control activities in the survey activities questionnaire.

International quality control program

Compliance with internationally standardized procedures and guidelines is highly important to achieve the overarching goal of high-quality data, which is internationally comparable. The international quality control program, which was conducted during the ICILS main survey, was designed and carried out by IEA and its purpose was to monitor data collection activities for ICILS 2018 on a national level. The program was essential to ensure that participating countries took a rigorous, standardized approach to data collection. For this purpose, IEA appointed IQOs for each participating country. The IQOs' main task was to document data collection activities and determine whether the ICILS assessment was administered in compliance with the standardized procedures in their respective countries.

Selecting and training international quality observers (IQOs)

IQOs are independent experts who have experience with school environments and ideally the ICT context, reside in the participating country, and have no affiliation with the national study centers. In order to facilitate the recruitment process, IEA asked NRCs to nominate two candidates to serve as IQOs for ICILS 2018 in their respective countries, which included sending the candidates' CVs and a short explanation why they think the nominees would be suitable for the position to IEA. Potential candidates had to be familiar with school environments or the day-to-day operations of schools and needed to be ICT literate. Additionally, they needed to be fluent in both the administered language(s) in the target country and English. IQOs should also not have any professional or personal affiliation to the national study center. Potential candidates were, for instance, school inspectors, relevant ministry officials, retired school teachers, or school principals. After carefully reviewing the CVs and considering the NRC's recommendation, IEA selected the most suitable candidate for the IQO position in every country.

Following the selection process, IEA invited the selected candidates to Amsterdam for a one and a half day in-person training. During the training, they were familiarized with the content of the study, the general procedures, and their tasks and responsibilities as IQOs. They were also informed about the possibility of hiring IQO assistants to reduce their workload. This was particularly advised in the case of large countries in order to cover different regions, states, or territories of the country or in case of a very short data collection window. If IQOs decided to appoint an assistant, they were solely responsible for their training, communication, payment, and any other organizational tasks. However, IQOs needed to inform IEA about their assistants and were required to submit confidentiality agreements signed by the assistants. In addition to the training, IQOs received the following materials, which they needed to accustom themselves with the ICILS 2018 procedures, their responsibilities, and to complete their documentation:

- *International Quality Observer Manual*
- International versions of the principal, teacher, and ICT coordinator questionnaires
- National adaptation form (NAF) template
- International version of the *School Coordinator Manual*
- International version of the *Test Administrator Manual*
- School visit travel form
- School visit tracking form
- Administration observation record
- Translation verification report
- Checklist for collecting materials from the NRC
- Confidentiality agreement to be signed by IQO assistants

In addition to the information provided to IQOs during the training seminar, the above-listed materials contained all necessary instructions needed to serve as IQO for ICILS 2018. After the training seminar and during the IQOs' fieldwork, IEA remained in close contact with the IQOs and provided further assistance and instructions if required, especially in case of unforeseen circumstances.

Overview of IQOs' responsibilities

The IQOs tasks and responsibilities were outlined and described in great detail in the IQO manual and conveyed during the training seminar. Their main responsibility was to monitor the data collection activities in their countries, document their findings, and report them to IEA. More precisely, IQOs' tasks covered the following:

- Familiarizing themselves with the ICILS content and context
- Visiting the NRC and collecting national instruments and other materials
- Selecting 15 schools for international quality control
- Contacting the schools selected for international quality control and arranging visits
- Observing an ICILS testing session in each selected school
- Interviewing the school coordinator and test administrator of each selected school
- Completing the translation verification report
- Completing the documentation and reporting to IEA

Visiting the national research coordinator (NRC)

Before the start of the data collection period for the ICILS 2018 main survey, IQOs were required to visit the NRC in their country. During this visit, IQOs had to select a sub-sample of schools for the observation of the student assessments in collaboration with the NRC. In addition, they had to collect national materials that were needed to conduct the school visits and a set of the national study instruments. NRCs provided IQOs with the following materials:

- USB flash drive containing the national version(s) of the student test and questionnaire
- Printed or digital version of the ICT coordinator, principal, and teacher questionnaire
- Printed copies of the national version(s) of the *School Coordinator Manual* for each administered language
- Printed copies of the national version(s) of the *Test Administrator Manual* for each administered language
- Student listing and tracking forms for each selected school
- Teacher listing and tracking forms for each selected school
- Contact information for the selected schools

The listed forms and documents were needed as a reference during the testing session observations, for the interviews with the school coordinators and test administrators, and to complete the IQOs' documentation. After IQOs completed their tasks and fulfilled their contract, they were asked to send all materials that they received from the NRC to IEA.

The task for determining the subsample of schools for international quality control included selecting 15 schools, plus three extra schools as replacements in case the IQO had difficulties contacting the initially selected schools or other unforeseen circumstances. Ideally and to the extent possible, the sampled schools were chosen following a random selection process. However, a random selection could be subject to a number of practical constraints. Therefore, IQOs were allowed to exclude schools outside of a reachable driving distance from where they or their assistants were residing. Another practical constraint was that the school should not already be selected for participation in the national quality control program (see the *Survey activities questionnaire* section below for more information about this program). Despite these constraints, IQOs were asked to attempt to visit schools in different areas or regions of their country in order to ensure a decent geographical coverage. In the case of very large countries, IQOs were highly advised to appoint assistants for this purpose. Following the selection of the schools to be visited for international quality control, IQOs needed to send their selection to IEA for approval. For this

purpose, IEA mainly assessed the number of selected schools, geographical coverage, and financial implications. Due to the country size of Luxembourg and its smaller sample size of schools, the study consortium approved that the IQO in this country would visit only 10 schools. For Germany, the IQO visited 20 schools in total; 15 for Germany as a whole and additional five schools for the benchmarking region North Rhine-Westphalia. All together, IQOs visited the agreed upon number of schools, which were in total 195 schools in 14 educational systems.

Testing session observations

The main responsibility of IQOs was to conduct an on-site observation of an ICILS testing session in each of the selected schools. For this purpose, IQOs contacted the school coordinators for every selected school prior to the beginning of the main survey data collection to obtain information about the testing day and to arrange the visit. In the case of a very short testing window in the respective country or colliding testing days, IQOs could either appoint assistants or consider using one of the replacement schools. In preparation for the observation of the testing session, IQOs needed to prepare print-outs or have an electronic device where they could access the administration observation record, and the student and teacher listing and tracking forms for the respective schools. They also had to familiarize themselves with the *Test Administrator Manual* and *School Coordinator Manual*. While IQOs 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 administration observation record after every school visit. The online version of the administration observation record was accessible through the IEA Online Survey System and all information needed to be entered in English.

While observing, IQOs were asked to use the administration observation record to structure their observation of the testing session and document their findings. The administration observation record included multiple choice and open ended questions, and consisted of four different content areas: (a) the ICILS administration arrangements and settings; (b) the ICILS administration process; (c) summary observations and general impressions of the ICILS administration; and (d) the interview with the school coordinator and test administrator. The following section presents the data derived from these administration observation records.

ICILS administration arrangement

The first section of the administration observation record asked about the assessment conditions and general preparations regarding the ICILS administration. It covered the setup of the room where the session took place and the arrangement of assessment materials. According to the answers in the administration observation record, around two thirds of the IQOs (67%) met the school coordinator prior to the preparation of the assessment room upon arriving at the schools. The preparation of the assessment room began about 60 minutes before the testing session started in around half of the schools (49%) and in approximately a third of the cases (36%) even earlier. IQOs also reported in most of the cases (91%) that the testing materials were safely stored and securely sealed when they arrived at the school.

Generally, the testing rooms seemed in order and well prepared in the majority of the visited schools. Table 9.1 shows detailed answers about the set-up of the testing rooms and preparations, which includes the seating space for students, the set-up of the work stations, and preparatory actions taken by the test administrator.

Table 9.1: Preparation of the testing room

Question	Response category (%)		
	Yes	No	Missing
Did the test administrator receive an adequate supply of USB flash drives? (if applicable)	99	1	–
Is there adequate seating space for students to avoid unwanted distractions while testing is in progress?	95	5	–
Is there adequate space for the test administrator to move around the room while the testing is in progress?	98	2	–
Does the test administrator have a watch or use another type of device to keep track of time while the testing is in progress?	97	3	–
Did the test administrator set up all workstations with each of them displaying the welcome screen prior to the students' arrival?	81	19	–
Is the test administrator ensuring that each student is sitting in front of the computer specially prepared for him or her?	93	6	1

Note: Percentages derived from a total of 195 responses.

During the ICILS administration

In the second section of the administration observation record, IQOs documented their observations of various aspects related to the administration of the student test and student questionnaire. This included observing and reporting on how accurately the test administrator followed the prescribed administration procedures and administration scripts. To complete this section in the administration observation record, IQOs had to refer to the administration scripts in the *Test Administrator Manual* and to the student tracking form. IQOs documented if test administrators followed the script, which was largely the case (see Table 9.2). If changes were made to the script, they were mainly of a minor nature. In these cases, the test administrators added information or revised instructions from the script slightly. Only in rare cases was content deleted, for example, if students were inattentive or lost concentration.

Table 9.2: Test administrator adherence to the administration script

Script type	Response category (%)			
	No changes	Minor changes	Major changes	Missing
ICILS assessment	68	28	4	–
Computational thinking modules*	76	23	1	–
Student questionnaire	76	20	3	2

Notes: Percentages derived from a total of 195 responses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.
* Only applicable in eight countries

Concerning the computers or tablets used for the ICILS administration, IQOs reported that in almost all cases (92%) the modules being displayed on the screens during the ICILS assessment corresponded with the information listed on the student tracking form for each student in the room. In almost all observed testing sessions (99%) exiting of the testing software of the student devices ran smoothly and the screens reverted back to the initial login screen.

As for students' compliance with the allocated time for the administration, IQOs reported that the majority of the students (92%) stopped working immediately after the allowable time period for the ICILS assessment ended. For countries that administered the computational thinking (CT) module, the majority of the students (96%) stopped working immediately after the allocated time period for the CT modules ended as well. In nine percent of the cases, there were students in the testing session who completed the assessment before the end of the allowable time period. Regarding the administration of the student questionnaire, approximately a third of the students (32%) asked for additional time to complete the questionnaire. In these cases students were given between one and 30 additional minutes to complete the student questionnaire (on average around seven minutes).

General impressions of the IQO

The third section of the administration observation record was dedicated to general impressions and observations of the IQO during the ICILS administration. In general, IQOs described the overall quality of the ICILS administration in about half of the cases (46%) as excellent, in around a third of the cases (35%) as very good, and in around an eighth of the cases (13%) as good. A few sessions were considered only of fair (5%) or poor (1%) quality. Most IQOs were also under the impression that test administrators familiarized themselves with the procedures and scripts prior to the test administration (95%).

Concerning the technical infrastructure at the schools, IQOs did not observe many technical issues (see Table 9.3). In some cases, problems when logging in, using the USB flash drives, or other malfunctions when using the ICILS assessment delivery system occurred. Other malfunctions included appearing error messages, sudden interruption or locking of a module, or frozen screens. IQOs reported on the necessity to change tablets, replace USB flash drives, or to restart computers in some testing sessions to resolve the problems that occurred. IQOs considered the actions taken by the test administrators to solve problems as efficient most of the time (95%).

Table 9.3: Technical problems experienced with procedures or devices

Procedures/devices	Response category (%)		
	Yes	No	Missing
Logging in	13	87	–
Using USB flash drives	13	72	15
Timer function	2	97	1
Other malfunctions when using the delivery system	21	78	1
Keyboard (switching languages)	3	94	3

Note: Percentages derived from a total of 195 responses.

Further, IQOs were under the impression that test administrators recorded students' attendance on the student tracking form correctly (100%). In total, there were only a few sessions with students who refused to take the assessment either prior to or during the administration (5%). IQOs reported that in about three quarters of the observed sessions (76%) students did not have particular problems with the ICILS administration. If there were issues, IQOs reported about students getting tired or losing concentration due to the length of the assessment. Furthermore, it appears that some students had difficulty understanding certain exercises. Overall, IQOs considered the students in around two thirds of the observed testing sessions as extremely orderly and cooperative (68%) and in a bit less than a third of the sessions (28%) as at least moderately orderly and cooperative. In most of the sessions (87%) IQOs did not observe students attempting to cheat or students who did not pay attention. In the case of non-cooperative and disorderly students, IQOs reported that most of the test administrators made an effort to control the students and the situation (84%).

Interview with school coordinators and test administrators

The final section of the administration observation record was dedicated to interviews with the school coordinator and test administrator. The purpose of these interviews was to obtain their evaluation of the ICILS administration, gather suggestions for improvement, and acquire additional background information.

School coordinators were mainly internal school staff. Fifty-one percent were school principals or other members of school management, 37 percent were teachers in the school where the assessment took place, and nine percent were other school staff members. In contrast, when it comes to test administrators, about a third of them were internal to the school where the assessment took place and about two thirds were external. More precisely, they were principals (8%), teachers of ICT (10%), teachers of another subject (13%), other school staff (6%), ICILS national center staff (17%), or they had an external position (46%).

When asking about the overall quality of the ICILS administration, 77 percent of all school coordinators said that it went very well and without problems and 16 percent answered that it went satisfactory and only with a few problems. If there were problems, school coordinators sometimes described difficulties with organizing and arranging the administration. In some schools, there were technical issues or problems with using technology. School coordinators rated the general attitude of other school staff members towards the ICILS assessment as positive (58%) or neutral (32%) in most of the cases. Furthermore, school coordinators reported to IQOs that approximately half of the students (54%) received some sort of special instructions, motivational talks, or incentives to prepare them for the assessment.

With regard to the *School Coordinator Manual*, the majority of the school coordinators (89%) felt that the manual worked well and did not need improvement. However, some suggestions for improvement were given and included shortening the manual, a better structure of the manual, or clearer instructions in certain areas. In addition to the *School Coordinator Manual*, around half of the school coordinators (47%) received some sort of additional training, instructions, motivational talks, or incentives from the national centers. Regarding the *Test Administrator Manual*, three quarters of the test administrators (75%) felt the manual worked well and did not require improvement. Some test administrators suggested improvements that included more specific explanations for certain aspects of the assessments, clearer instructions, or reducing redundancy.

IQOs asked school coordinators about various forms used during the ICILS administration and if the information on these forms was correct. Table 9.4 shows that teacher listing forms and student listing forms contained correct information in the majority of the schools, and that the teacher questionnaires or cover letters were distributed according to the teacher tracking form in almost all cases.

Table 9.4: Teacher and student listing forms and teacher tracking forms used for the assessment

Question	Response category (%)		
	Yes	No	Missing
Did the teacher listing form include all eligible teachers as listed in the school timetable for the target grade?	93	5	2
Were all students participating in the ICILS assessment listed on the student listing form?	93	4	3
Were the teacher questionnaires/cover letters distributed according to the teacher tracking form?	85	9	7

Notes: Percentages derived from a total of 195 responses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

Translation verification report

All national study instruments underwent a thorough translation verification process before administration (see Chapter 5 for more information). The purpose of translation verification was to ensure that the language remains faithful and equivalent in meaning as intended by the international source version of the instruments. Upon completion of translation verification, the instruments containing translation verifier comments and suggestions were released back to the NRC. The NRC reviewed the verifier's feedback and had the final decision on whether or not to adopt the recommendations. The consortium requested all NRCs to respond to the verifier's feedback and to document whether they agree or disagree with the verifier's suggestion or if they want to modify the translation further.

The IQO's task was to review the final national study instruments and the NRC's response to the translation verification feedback. They had to check whether or not the NRC adopted the verifier's suggestions and whether or not they documented their actions correctly. For this purpose, IQOs referred to the final set of teacher, principal, ICT coordinator, and student questionnaires, and the student test modules, which they received from the NRC. In addition, IEA provided IQOs with translation exports from the translation system, which included the NRC's response to the verifier's comments. IQOs had to compare the translations after translation verification with the final translations and confirm if the NRC's response corresponded with the applied changes. In addition, IQOs were asked to give their opinion on the overall quality of the verifier's work and the overall use of the verifier's feedback by the NRC. IQOs had to document their findings in the translation export and the translation verification report.

In addition to the review of the ICILS 2018 study instruments, IQOs had to report on the appropriateness of the national adaptation of the *School Coordinator Manual* and *Test Administrator Manual*. National study centers had to adapt these manuals to their national context. The national versions did not undergo any verification process performed by the consortium. The IQO's task was to review the national versions of these manuals and document if they were consistent with the international templates. For the *School Coordinator Manual*, IQOs checked the alignment of the sections that described the role of the school coordinator, preparing for within-school sampling of students and teachers, preparing for the administration, administering the questionnaires, and quality control during the main survey. For the *Test Administrator Manual*, IQOs reported on the alignment of sections about the role of the test administrator, conducting the assessment, administering the test, and troubleshooting. IQOs also documented if additions were made to the national version of the manuals.

Survey activities questionnaire

The survey activities questionnaire (SAQ) is a comprehensive questionnaire that covers different areas of the implementation of the ICILS 2018 main survey procedures. National centers of all participating educational systems[1] gave their feedback on the general approach, standardized procedures, and materials provided by the consortium (e.g., manuals, software, forms) for the ICILS 2018 main survey. They also reported on difficulties they experienced and provided suggestions for improvements for future ICILS cycles. The SAQ addressed the following areas of interest:

- Sampling
- Contacting schools and recruiting school coordinators
- Implementing and documenting national adaptations using the NAF
- Translating instruments using the IEA Translation System and the AssessmentMaster (AM) system from RM Results, AM Designer.

1 Germany and North Rhine-Westphalia collaborated for filling in the SAQ. Therefore, there were 13 country entries in total.

- Assembling and preparing the ICILS materials for administration
- Preparing online adult questionnaires using the IEA Translation System
- Administering ICILS
- National quality control activities
- Scoring open-ended response items
- Coding occupation data using the IEA Coding Expert software
- Entering data manually and submitting data
- Time required for survey activities
- Other experiences

Overall, the feedback provided by the NRCs was considered valuable, fruitful, and helped the consortium to gain a better perspective on the survey activities at the national level. The following sections summarize the most important results and insights from the SAQ.

Sampling

With regard to collaboration with the IEA sampling team, all NRCs reported that they felt well-supported for all sampling related matters. They further informed the consortium that they did not find it difficult to adapt the international sampling design to their national specifications. Only one country considered this as somewhat difficult. The same is the case for creating a sampling frame, which lists all eligible schools. Twelve out of 13 national centers did not find this task difficult at all. NRCs further considered the Survey Operations Procedures Unit 1 (Sampling Schools) as sufficiently describing the procedures when it comes to defining and identifying the target populations of the survey and developing national sampling plans. National study centers reported that they had no or only some difficulties to select the student and teacher samples with the IEA Windows Within-School Sampling Software.

Contacting schools and recruiting school coordinators

National study centers contacted the sampled schools in multiple ways. Eleven countries contacted the schools by phone, six by regular mail, eleven by email, and five in other ways including websites or personal visits. For contacting the schools, nine national study centers used letters based on other national projects, one used example letters provided in the Survey Operation Procedures Unit 2 (Working with Schools), and three used other kinds of letters.

The SAQ further asked if schools' participation in ICILS 2018 was compulsory in the participating countries. Seven countries reported that participation was not compulsory in any school. In four countries participation was compulsory for all schools and in two countries it was compulsory for some of the schools, for instance, due to different regulations in federal states. Nine out of 13 NRCs reported having difficulties in convincing schools to participate. Reasons for this included schools' participation in other international studies, ongoing school reforms, or busy school schedules.

National study centers trained school coordinators in different ways and sometimes combined training methods. The training methods included formal in-person training sessions, instructions via telephone, emails, videos, and/or conventional written instructions. Other methods included webinars or visiting schools in person. Table 9.5 shows more detailed information about school coordinator trainings as well as difficulties school coordinators reported with understanding certain aspects of the ICILS administration.

Table 9.5: School coordinator information

Question	Response category (%)	
How did you train the school coordinators?	Yes	
• Formal training session	5	
• Through telephone, email, or video-link	4	
• Written instructions	8	
• Other	5	
Question	**Yes**	**No**
Did the school coordinator report difficulties with understanding any of the following aspects of ICILS administration?*		
• Identifying eligible teachers and/or students	4	8
• The necessity for listing all target grade teachers	3	9
• The necessity for listing all target grade students	4	8
• Flagging students to be excluded prior to school sampling	4	8
• The rationale for sampling students from the target grade across classrooms	1	11
• The process for running the USB compatibility tests in schools	5	7
• The steps to set up computers in schools to support successful test administration	3	9
• The test administration procedures	1	11

Notes: Percentages derived from a total of 13 responses.
* One country did not answer this question

Implementing and documenting national adaptations and translating instruments

The SAQ asked national centers about their experiences with the adaptation of the international versions of the ICILS 2018 study instruments. None of the countries reported difficulties adapting the student test modules and the school principal questionnaire. A few countries had difficulties translating the student questionnaire (2), the teacher questionnaire (1), and the ICT coordinator questionnaire (1). When documenting national adaptations to the NAFs, almost all countries reported not having difficulties (11).

Around half of the countries (6) did not report difficulties with translating the student instruments into national languages using AM Designer, which was provided by the consortium for translating the student test and questionnaire. The rest of the countries (7) experienced some difficulties with the translation process. These difficulties included technical problems when using the translation system and general functionality. Regarding the translation of the ICILS principal, teacher, and ICT coordinator questionnaires using the IEA Translation System, the majority of the countries did not experience any difficulties (10). If there were difficulties, countries reported issues with the general usability and user-friendliness of the system. Translation verification feedback provided by IEA was considered by almost all countries (12) as very useful or at least somewhat useful. In general, most of the national study centers did not report major problems regarding the adaptation verification (12) and translation verification (11) processes.

Preparing the ICILS materials for administration

National centers were asked to share their experiences with the preparations of the ICILS materials for administration. A few countries (4) reported having difficulties during the layout verification process of the ICILS delivery system for the student instruments. These difficulties were mainly due

to the functionality and usage of the system. The majority of the countries (12) did not report any major problems when it came to the verification process for the online questionnaires. Most of the countries did not experience major difficulties when downloading and replicating the ICILS delivery system to the USB flash drives, which included downloading the test file image (13), extracting the test file to a USB (12), and creating copies of the USB flash drives for use in the testing (11).

Administering ICILS

Some countries (4) reported problems concerning the web-application that occurred during the administration of the online questionnaires. Additionally, five countries observed some issues with the login procedure during the administration of the online questionnaires. These included issues with accessing the online questionnaire or that respondent identification (ID) and passwords did not work properly. Problems during the administration of the ICILS assessment included malfunctions of the delivery system (e.g., freezing of the screen), which in some cases led to the rebooting of the delivery system or other technical issues (e.g., corruption of USB flash drives, issues with the firewall).

Concerning the participation of the ICILS populations, some national study centers experienced difficulties in achieving high participation rates. Six countries reported problems with student participation, mainly due to difficulties with receiving parents' consent or student refusal. Eight countries experienced problems with teachers' participation. Reasons included teachers' workload, lacking motivation, or missing support of teacher unions. Four countries reported problems with ICT coordinators' and principals' participation rates.

Scoring open-ended response items

Scoring activities for open-ended response items in the participating countries were mainly performed by national center staff (3), teachers or professional educators (9), or university students (5). On average, 12 scorers were used per country. A few countries (5) reported that their scorers had difficulties using the scoring system for training and scoring the student work, which included problems with accessing the software, speed of the application, or displaying pictures appropriately. Further, six countries reported difficulties with the reports for reliability scoring.

Entering and coding occupation data

For coding occupations, most countries used mainly their own staff members (9). In addition, a few countries used experts from external organizations (3) or university students (1). On average, countries used four coders for coding of the occupation data. Most of the national study centers (12) made use of the International Standard Classification of Occupations (ISCO-08) scheme for coding occupations and used either existing translations or their own translations of the scheme. None of the countries reported problems with converting the national coding scheme to ISCO-08. Seven countries reported having difficulties with coding the student responses and one country about working with the IEA Coding Expert software.

Entering data manually and submitting data

Entering data manually was only applicable for three countries, as all other countries administered the teacher, principal, and ICT coordinator questionnaires solely online. One of these three countries used the IEA Data Management Expert (IEA DME) software to enter the derived data from the questionnaires.

National quality control activities

In addition to the international quality control program, quality control at the national level also occurred during the data collection. National study centers were responsible for conducting a national quality control program, for which the composition was at the discretion of the national study center. However, the consortium provided recommendations and guidelines in the form of a *National Quality Observer Manual* to NRCs. The recommendations included recruiting and

training NQOs, visiting 10 percent (or a minimum of 15) of the sampled schools to observe a testing session, and to interview the school coordinator and test administrator in each school. The manual included instructions for NQOs and a template ICILS administration observation record. However, countries could decide whether they would use the manual and could adjust it where necessary.

All 13 participants reported that they conducted national quality control. However, specific national quality control activities varied from country to country. Each country appointed one or more NQOs, who were mainly staff members of the national center, external experts, or school inspectors. NQOs were trained in multiple ways including formal training sessions (8), instructions via telephone, email, or videos (2), and written instructions (6). A minimum of six schools was visited in the course of the national quality control program in each country. Two countries reported conducting some sort of national quality control activities in all sampled schools, which included, for example, regular check-in calls with the test administrators. In one of these two countries, every testing session was observed on-site by a staff member of the national study center. Most national centers (12) reported that the visited schools were located in different regions of the country. The majority of the countries (11) made use of at least parts of the *National Quality Observer Manual* provided by the consortium. Adjustments to the manual included reducing the length of the document or adapting it to the national context. The manual was also used as a basis for producing training materials. Issues reported by NQOs included technical issues such as freezing or crashing of the software or making use of the USB compatibility check. NQOs further reported on minor procedural deviations, which included non-adherence to the allocated time for the testing session. They also provided feedback regarding the teacher and student tracking forms used during the administration and made suggestions for improvements.

Summary

The ICILS consortium, in cooperation with the participating countries, developed and coordinated a range of quality assurance measures in order to monitor the quality of the study administration. This chapter focused on two quality assurance activities that occurred during the main survey for ICILS 2018, namely the international quality control program and the SAQ. The information presented focused on the structure and general approach of these two components as well as the most important results.

The IQOs were appointed to observe 15 testing sessions of the ICILS student assessment in each participating country and interview the school coordinator and test administrator. Their main task was to monitor compliance with the internationally standardized procedures and report their findings to IEA. The SAQ was completed by all national study centers and shed light on different areas of the implementation of the ICILS 2018 main survey procedures, such as sampling, study preparation activities, administration of ICILS 2018, or data processing. NRCs provided their feedback on the general approach to ICILS 2018, the procedures, and support materials provided by the consortium.

Taken together, these activities form an important source of information about the implementation of different steps and processes of ICILS 2018, taking into account different perspectives from the national and international levels.

CHAPTER 10:

Data management and creation of the ICILS 2018 database

Ekaterina Mikheeva and Sebastian Meyer

Introduction

This chapter describes the procedures for checking the ICILS 2018 data and database creation that were implemented by IEA, the ICILS international study center (ISC) at ACER, and the national centers of the participating countries.

Preparing the ICILS 2018 international database and ensuring its integrity was a complex endeavor requiring extensive collaboration between IEA, the ISC, and the national centers. Once the countries had created their data files and submitted them to IEA, 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 codebooks and documentation appropriately reflected national adaptations to questionnaires;
- All variables used for international comparisons were comparable across countries (after harmonization where necessary); and
- All institutions involved in this process applied quality control measures throughout, in order to assure the quality and accuracy of the ICILS 2018 data.

Data sources

Computer-based student assessment

As a computer-based assessment, ICILS 2018 exclusively generated electronic data from the student assessment along with paper or electronic tracking information. In general, one version of the student assessment software existed per country. It included all test modules and components, and supported all assessment languages.

Each student was assigned an instrument version that contained two of the five 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. As the default administration method for ICILS 2018, the national centers provided schools with one USB stick per student (plus spare sticks in case of technical failure). If the computers at a school were deemed unsuitable for administering the student assessment, laptop computers were provided. The national 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 the administrators to indicate student participation for subsequent verification.

In schools where the default administration method was utilized, data from the student assessment were stored on the individual USB sticks used to deliver the assessment. If the server method was utilized, then the student data were stored on the server laptop that was used to administer the assessment. At the time of running the student assessment, data were saved in a long data format. Each form of interaction performed by the student was saved as an "event" with a unique timestamp. After the assessment, the national centers used an upload tool to upload the data to a central international server. This tool was provided either on the USB sticks or on the server computer for those schools that administered the assessment through the laptop server.

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 school and teacher questionnaires

ICILS 2018 offered online collection of school and teacher questionnaire data. All participating countries adopted the online option as a default data-collection mode for all respondents (that is, school principals, ICT coordinators, and teachers). National centers had to ensure that individual respondents who refused to participate in the online mode or who did not have access to the required infrastructure for online participation were provided with a paper questionnaire, thereby ruling out unit nonresponse as a result of a forced administration mode.

To ensure confidentiality, national 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 checksum provided from the IEA Windows Within-School Sampling Software (IEA WinW3S), meaning that the identity validation step conducted at the national centers for paper-based questionnaires occurred when the respondents logged into the survey.

As respondents completed their online questionnaires, their data were automatically stored on the central international server. Data for each country-language combination were stored in a separate table on the server. The different language versions within countries were then merged (at IEA) with the data collected as part of the within-school sampling process.

Potential sources of error originating from the use of the two parallel modes had to be kept to the absolute minimum to ensure uniform and comparable conditions across modes and countries. To achieve this, ICILS 2018 questionnaires in both modes were self-administered, had identical contents and comparable layout and appearance, and required the data collection for both modes to take place over the same period of time.

Data entry and verification of paper questionnaires

Data entry

Each national center was responsible for transcribing the information from paper-based principal, ICT coordinator, and teacher questionnaires into computer data files using the IEA Data Management Expert (IEA DME) software.

National centers entered responses from the paper questionnaires into data files created from an internationally predefined codebook. The codebook contained information about the names, lengths, labels, valid ranges (for continuous measures or counts) or valid values (for nominal or ordinal questions), and missing codes for each variable in each of the three nonstudent questionnaire types.

IEA provided countries who needed to manually enter data from paper-based questionnaires with a codebook that reflected the nationally adapted and internationally verified questionnaire structure. National codebooks were exported directly from the online system to ensure consistency between online questionnaires and the data entry codebook.

In general, national centers were instructed to discard any questionnaires that were unused or that were returned completely empty, and to enter any questionnaire that contained at least one valid response. To ensure consistency across participating countries, the basic rule for data entry in the IEA DME required data to be entered "as is" without any interpretation, correction, truncation,

imputation, or cleaning. Resolution of any inconsistencies remaining after this data entry stage would occur during data cleaning (see below).

The rules for data entry included the following:

- Responses to categorical questions to be generally coded as "1" if the first option was used, "2" if the second option was marked, and so on.
- Responses to "mark-all-that-apply" questions to be coded as either "1" (marked) or "9" (not marked/omitted).
- Responses to numerical or scale questions (e.g., school enrolment) to be entered "as is," that is, without any correction or truncation, even if the value is outside the originally expected range (e.g., if an ICT coordinator reports more than 1000 computers available to students in the school).
- Likewise, responses to filter questions and filter-dependent questions to be entered exactly as filled in by the respondent, even if the information provided is logically inconsistent.
- If responses were not given at all, not given in the expected format, ambiguous, or in any other way conflicting (e.g., selection of two options in a multiple-choice question), the corresponding variable was to be coded as "omitted or invalid."

Data entered with the IEA DME were automatically validated. First, the entered respondent ID was validated with a five-digit code—the checksum (generated by the IEA WinW3S). A mistype in either the ID or the checksum resulted in an error message that prompted the data-entry person to check the entered values. The data-verification module of the IEA DME also enabled identification of a range of problems such as inconsistencies in ID codes and out-of-range or otherwise invalid codes. Individuals entering the data had to resolve problems or confirm potential problems before they could resume data entry.

Double-data entry

To check the reliability of the data entry within respective participating countries, national centers were required to have all of the paper-based principal, ICT coordinator, and teacher questionnaires entered by two different staff members. IEA recommended that national centers begin the double-data entry process as early as possible during the data capture period in order to identify possible systematic, incidental misunderstandings or mishandlings of data entry rules and to initiate appropriate remedial actions, for example, retraining staff. Those entering the data were required to resolve identified discrepancies between the first and second data entries by consulting the original questionnaire and applying the international rules in a uniform way.

While it was desirable that each and every discrepancy be resolved before submission of the complete dataset, the acceptable level of disagreement between the first and second data entry was established at one percent or less; any value above this level required complete re-entry of the data. This restriction guaranteed that the margin of error observed for processed data remained well below the required threshold.

The level of disagreement between the first and second data entry was evaluated by IEA. Data for those countries who had administered paper-based questionnaires and submitted an IEA DME database showed no differences between the main files and the files created for the purpose of double-data entry.

Data verification at the national centers

Before sending the data to IEA for further processing, national centers were to carry out mandatory validation and verification steps on all entered data and apply corrections as necessary. The corresponding routines were included in the IEA DME software, which automatically and systematically checked data files for duplicate ID codes and data outside the defined valid ranges or value schemes. Data managers reviewed the corresponding reports, resolved any inconsistencies,

and (where possible) corrected problems by looking up the original survey questionnaires. Data managers also verified that all returned non-empty questionnaires had definitely been entered. They also checked that the availability of data corresponded to the participation indicator variables and entries on the tracking forms and as entered in the IEA WinW3S.

In addition to submitting the data files described above, national centers provided IEA with detailed data documentation, including hard copies or electronic scans of all original student and teacher tracking forms and a report on data-capture activities collected as part of the online survey activities questionnaire. IEA 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 being entered, each national center used the information from the teacher tracking forms to verify the completeness of the materials. Participation information (e.g., whether the concerned teacher had left the school permanently between the time of sampling and the time of administration) was entered via the IEA WinW3S.

This process was also supported by the option in the IEA 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 IEA DME, the database for online respondents, and the uploaded student data on the central international server. Data managers were requested to resolve these problems before final data submission to IEA. If inconsistencies had to remain or the national center could not solve them, IEA asked the center to provide documentation on these problems. IEA used this documentation when processing the data at a later stage.

Confirming the integrity of the national databases

Overview

As described earlier in this chapter, national centers in each participating country were responsible for entering their national ICILS 2018 data into the appropriate data files and submitting these files to IEA. Furthermore, the data from the online questionnaires were automatically stored on a central international server. IEA then subjected these data to a comprehensive process of checking and editing. To facilitate the data cleaning process, IEA asked the national 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, the national versions of test booklets and questionnaires, as well as information from the survey activities questionnaire (see details in Chapter 6). National centers also submitted their final national adaptation forms in order to provide and confirm complete documentation on all national adaptations. In addition, national centers were asked to provide documentation on all changes or edits applied to the data prior to submission, as well as any verified findings that could remain.

Ensuring the integrity of the international database required close cooperation between the international and national institutions involved in ICILS 2018. After each country had submitted its data and required documentation, IEA, in collaboration with the national research coordinators (NRCs), conducted a four-step cleaning procedure upon the submitted data and documentation:

(1) Documentation and structure check;

(2) ID variable cleaning;

(3) Linkage cleaning; and

(4) Background cleaning (resolving inconsistencies in questionnaire data).

The cleaning process was an iterative process. Numerous iterations of the four-step 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 centers, and all corrections made during the cleaning process were documented in a cleaning report produced for each country.

During the first step, IEA checked the data files provided by each country. In the following steps, they applied a set of over 120 cleaning rules to verify the validity and consistency of the data and documented any deviations from the international file structure.

Having completed this work, IEA staff sent queries to the national centers. These required the centers to either confirm IEA's proposed data-editing actions or provide additional information to resolve inconsistencies. After all modifications had been applied, IEA 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 files had been checked, IEA provided national centers with univariate statistics at the national and international levels. This material enabled national 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 the comparability of the data across countries. 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 international team 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, IEA sent data to the ISC, which then produced and subsequently reviewed the basic item statistics. At the same time, IEA produced data files containing information on the participation status of schools, students, and teachers in each country's sample. IEA then used this information, together with data captured by the software designed to standardize operations and tasks, 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 ISC estimated CIL performance and CT scores as well as questionnaire indices for students, teachers, and schools (see Chapters 11 and 12 for scaling methods and procedures). On completing their verification of the sampling weights and scale scores, the ISC sent these derived variables to IEA for inclusion in the international database and for distribution to the national centers.

Data cleaning quality control

Because ICILS 2018 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 applying the programs to real datasets, IEA applied them to simulation datasets containing all possible problems and inconsistencies.

- *Registering all incoming data and documents in a specific database*: IEA 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*: IEA recorded these in the ICILS 2018 general cleaning documentation for the main survey.
- *Logging every "manual" correction to a country's data files in a recoding script*: Logging these changes, which only occurred occasionally, allowed IEA to undo changes or to redo the whole manual cleaning process at any later stage of the data-cleaning process.
- *Repeating, on completion of data cleaning for a country, all cleaning steps from the beginning*: This step allowed IEA to detect any problems that might have been inadvertently introduced during the data-cleaning process.
- *Working closely with national centers at various steps of the cleaning process*: IEA provided national centers with the processed data files and accompanying documentation and statistics so that center staff could thoroughly review and correct any identified inconsistencies.

IEA compared national adaptations recorded in the documentation for the national datasets against the structure of the submitted national data files. IEA then recorded any identified deviations from the international data structure in the national adaptation database and in the *ICILS 2018 user guide for the international database* (Mikheeva and Meyer, 2020). Whenever possible, IEA recoded national deviations to ensure consistency with the international data structure. However, if international comparability could not be guaranteed, IEA removed the corresponding data from the international database.

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 school, teacher, and student information 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 following activities (summarized in Figure 10.1 and explained in the following subsections). IEA carried out all of these activities in close communication with the national centers.

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).

IEA began data cleaning by combining the tracking information and sampling information captured in the IEA WinW3S database with the student-level database containing the corresponding student survey instrument data. During this step, IEA staff also retrieved the data from the principal, ICT coordinator, and teacher questionnaires for both the online and paper administration modes. 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, IEA asked the national centers to complete national adaptation forms while they were adapting the international codebooks. Where necessary, IEA 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 Appendix 2 of the *ICILS 2018 user guide for the international database* (Mikheeva and Meyer, 2020).

At this stage, IEA discarded variables created purely for verification purposes during data entry, and made provision for new variables necessary for analysis and reporting. These included reporting variables, derived variables, sampling weights, and scale scores.

Once IEA had ensured that each data file matched the international format, they applied a series of standard data cleaning rules for further processing. Processing during this step employed software developed by IEA that could identify and correct inconsistencies in the data. Each potential problem flagged at this stage was identified by a unique problem number, described and recorded in a database alongside the specific action taken by the cleaning program or IEA in relation to the problem.

IEA referred problems that could not be rectified automatically to the responsible NRC so that the national centers could check the original data collection instruments and tracking forms to trace the source of these errors. Wherever possible, IEA suggested a remedy and asked the national centers to either accept or propose an alternative. If a national center could not resolve problems through verification of the instruments or forms, IEA applied a general cleaning rule to the files to rectify this error. When all automatic updates had been applied, IEA ran individual recodings in the data to directly apply any remaining corrections to the data files.

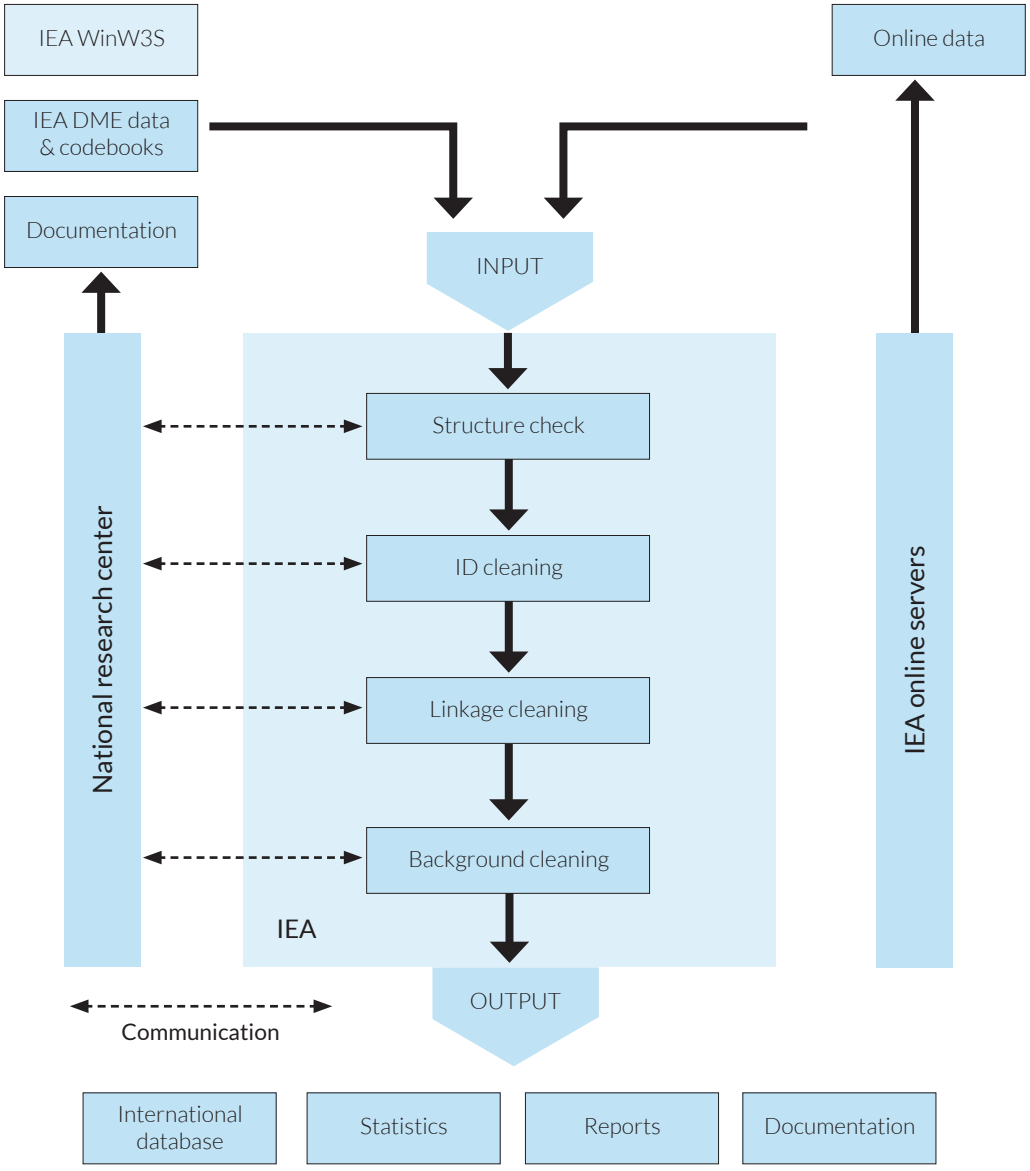
Cleaning identification (ID) variables

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 2018 database shared the same ID number and contained exactly the same data, IEA deleted one of the records and kept the other one in the database. If both records contained different data and IEA found it impossible to identify which record contained the "true data," they removed both records from the database. IEA 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. This step in data cleaning included the preparation of the student test records provided to IEA by RM Results. 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 extreme cases, students from an entire school had the same student ID.

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, re-identified, 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 in order 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 IEA with confirmation from national centers for individual records. Further details on cleaning of multiple records are reflected in Figure 10.1 below.

Figure 10.1: Overview of data processing at IEA



As mentioned earlier, IEA conducted all cleaning procedures in close cooperation with the national centers. After national centers had cleaned the ID variables, the clean databases with information about student participation and exclusion were used by the IEA sampling section to calculate students' participation rates, exclusion rates, adjudication flags, and student sampling weights (see Chapter 7 for details).

Checking linkages

In ICILS 2018, data about students, their schools, and teachers appeared in a number of 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 (Table 8.1) 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 in questionnaire data

The amount of inconsistent and implausible responses in questionnaire data files varied considerably among countries and none were completely free of inconsistent responses. IEA determined the treatment of inconsistent responses on a question-by-question basis, using all available documentation to make an informed decision. IEA also checked all questionnaire data for consistency across the responses given.

For example, Question 3 in the principal questionnaire asked for the total school enrolment (number of boys and number of girls, respectively) in all grades, while Question 4 asked for the enrolment in the target grade only. Clearly, the number given as a response to Question 4 could not exceed the number provided in Question 3. Similarly, it was not possible for the sum of all full-time teachers and part-time teachers, as asked in Question 6, to equal zero. In another example, Question 7 of the ICT coordinator questionnaire 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.

IEA flagged inconsistencies of this kind and then asked the national centers to review. IEA 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. IEA 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."

IEA also applied what is known as a split variable check to questions where the answer was coded into several variables. For example, Question 26 in the student questionnaire asked students: "At school, have you learned about the importance of the following topics?" Student 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, students 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 students had given affirmative responses in the other categories.

Resolving inconsistent tracking and questionnaire information

Two different sets of ICILS 2018 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 choose between five age ranges. Year of birth, which was indicated in the tracking forms, was then recoded into age groups and cross checked against the range indicated by the questionnaire responses. If gender and/or age range 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. However, for teacher birth year, tracking information was set to missing given that only an age range, not a specific year, was indicated.

Handling of missing data

Two types of entries were possible during the ICILS 2018 data capture: valid data values and missing data values. Missing data can be assigned a value of "omitted or invalid," or "not administered" during data capture. With the exception of the "not reached" missing codes assigned at ACER, IEA 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:

- *Omitted or invalid*: 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 schools 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 (this applied only to the 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. "Not reached" codes were derived as follows: an item received this coding if a student did not respond to any of the items following it within the same booklet[1] (i.e., the student did not complete any of the remaining test questions), if he or she did not respond to the item preceding it, and if he or she did not have sufficient time to finish a module in the booklet.

Checking the interim data products

Building the international database was an iterative process. Once IEA completed each major data processing step, it sent a new version of the data files to the national 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 centers received several versions of their data, *and their data only*, before release of the draft and final versions of the international databases. All interim data were made available in full to the ISC, whereas each participating country received only its own data.

IEA sent the first version of data and accompanying documentation to national centers on October 30, 2018. At this time, data for all countries who administered ICILS in the first half of 2018 were sent out. The data for countries who administered ICILS in the second half of 2018 received their data on January 16, 2019. This first version of each country dataset included the following data and documentation:

- School-, student-, and teacher-level SPSS and SAS data files;
- Univariate descriptive statistics for all variables in the data files;
- A cleaning report that included a list of structural and case-level findings;
- A recoding documentation for country-specific data edits applied by IEA; and

1 The term booklet is used here in reference to any possible combination of test modules.

- Cleaning documentation describing the initial cleaning procedures undertaken at IEA and describing the data files and statistics provided.

IEA provided the ISC with subsequent versions of the data and related documentation as soon as it had implemented feedback from national 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, IEA asked countries to review the documentation on adaptations to the national versions of their instruments and the related edits applied to the data files.

In May 2019 all national centers received the data from all other ICILS 2018 countries. This data version is called the draft international database since it roughly reflects the structure of the released databases. Most prominently and compared to earlier data send-outs this version included sampling weights and scales. All persons within the national centers needed to sign a confidentiality agreement assuring no sharing of any ICILS 2018 data products or information with respect to the results.

During the fifth NRC meeting in Jyväskylä, Finland in June 2019, NRCs had the opportunity to raise any further issues concerning their data that had not yet been raised.

In August 2019, IEA provided NRCs with an updated version of the draft international database. This version was necessary due to minor edits IEA and ACER were made aware of and which resulted from internal quality control procedures. The ISC used this version of the data to produce the updated, final tables for the international report.

The ICILS 2018 international database

The ICILS 2018 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;
- 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 2018 international database is provided in the *ICILS 2018 user guide for the international database* (Mikheeva and Meyer, 2020).

Summary

To achieve a high-quality database, ICILS 2018 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. IEA reviewed all national databases in cooperation with national 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 2018 database. The final data products included item statistics, national data files, and the international database accompanied by a user guide and supplementary information.

References

Mikheeva, E., & Meyer, S. (Eds.). (2020). *IEA International Computer and Information Literacy Study 2018 user guide for the international database*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/publications/user-guides/icils-2018-user-guide-international-database

CHAPTER 11:

Scaling procedures for ICILS 2018 test items

Louise Ockwell, Alex Daraganov, and Wolfram Schulz

Introduction

This chapter describes the procedures used to analyze and scale the ICILS 2018 test items that were administered to measure students' computer and information literacy (CIL) and computational thinking (CT). It covers the following topics:

- The scaling model used to analyze and scale the test items;
- Test coverage;
- Item dimensionality and local dependence;
- Assessment of item fit;
- Assessment of scorer reliabilities for open-ended items;
- Differential item functioning by gender;
- Review of cross-national measurement equivalence;
- International item adjudication;
- International item calibration and test reliability;
- International ability estimates (plausible values and weighted likelihood estimates); and
- Estimation of changes in students' CIL between 2013 and 2018.

The development of the CIL and CT test items is described in Chapter 2 and was guided by the ICILS 2018 assessment framework (see Fraillon et al. 2019).

The scaling model

Item response theory (IRT) scaling methodology was used to scale the test items.

For dichotomous items, we used the one-parameter (Rasch) model (Rasch 1960), which models the probability of selecting category 1 instead of 0 as:

$$P_i(\theta_n) = \frac{\exp(\theta_n - \delta_i)}{1+\exp(\theta_n - \delta_i)}$$

where $P_i(\theta_n)$ is the probability for person n to score 1 on item i, θ_n denotes the estimate of person n's location on the latent continuum (which in proficiency tests is commonly referred to as person *ability*) and δ_i is the estimated location of item i on the same latent continuum (which in proficiency tests is commonly referred to as item *difficulty*). For each item, item responses are modeled as a function of the latent trait θ_n.

In the case of items with more than two categories (in the CIL and CT assessments, items with more than one score point), this model can be generalized to the (Rasch) partial credit model (Masters and Wright 1997), which takes the form of:

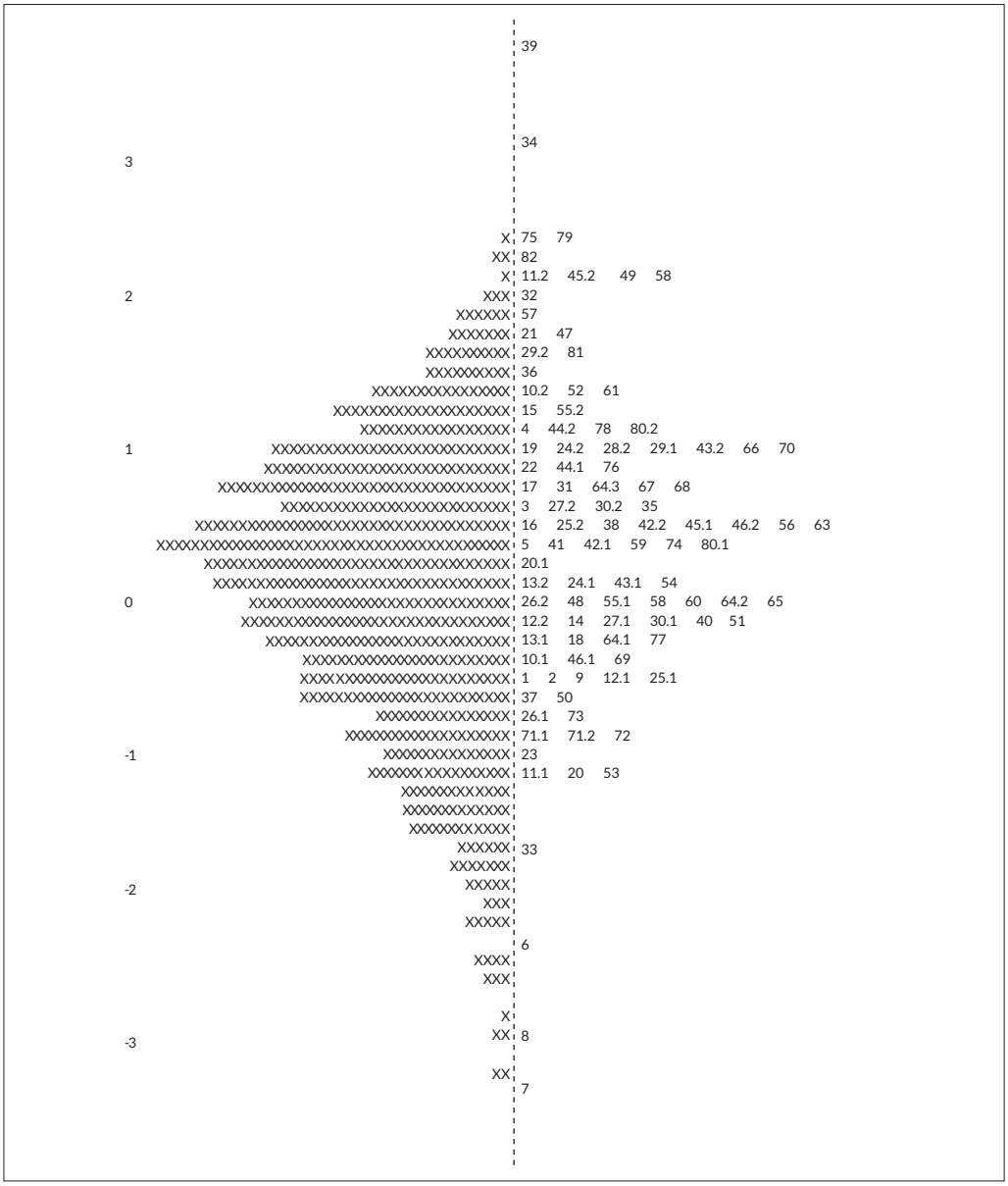
$$P_{x_i}(\theta_n) = \frac{\exp\sum_{j=0}^{x}(\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i}\exp\sum_{j=0}^{h}(\theta_n - \delta_{i+} + \tau_{ij})} \quad x_i = 0,1,\ldots,m_i$$

where $P_{x_i}(\theta_n)$ denotes the probability of person n scoring x on item i, θ_n denotes the estimate of the person n's location on the latent continuum, the item parameter δ_i denotes the estimated average location (across the categories) of the item on the latent continuum, and τ_{ij} provides an additional step parameter that denotes the distance between estimates of each category boundary and the estimated location of the item on the latent continuum. ACER ConQuest, Version 4.0 software (Adam et al. 2015) was used to scale the CIL and CT test items for ICILS 2018.

Test coverage and item dimensionality

When measuring cognitive abilities, it is important to use test items that cover the range of achievement found in the target population. First, we estimated the distribution of CIL and CT among ICILS 2018 students and the location of the corresponding item thresholds (with a response probability, *rp* = 0.5[1]; see Figures 11.1 and 11.2). Item thresholds were equal to item difficulties of dichotomous items. For partial credit items, a difficulty threshold was estimated for each score.[2]

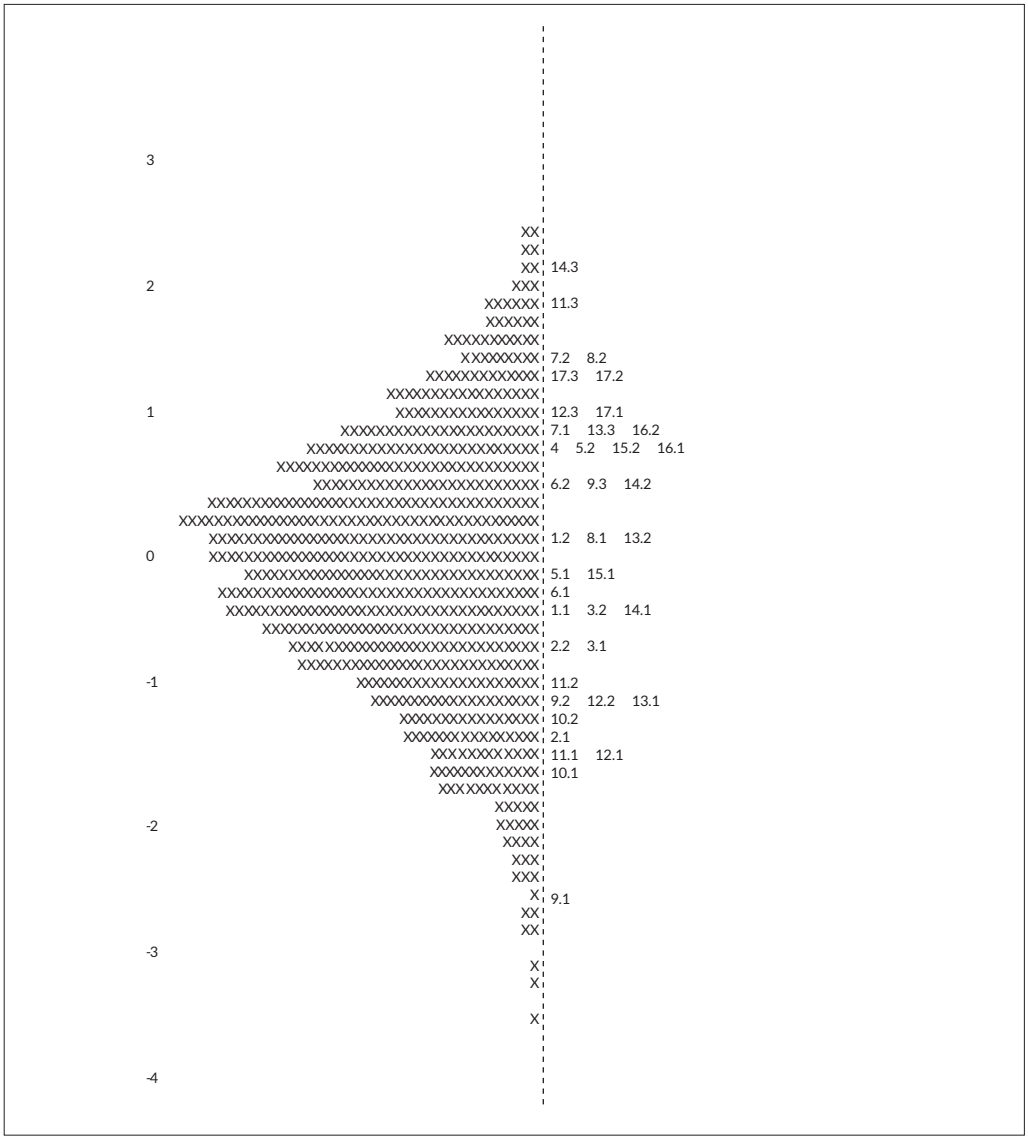
Figure 11.1: Mapping of CIL student abilities and item difficulties



1 This means that a respondent with the same score on the latent continuum as the item location parameter has a probability of 50 percent to give a correct response.

2 This "Thurstonian" threshold indicates for each item score the point on the location, where a respondent with the same latent trait score has a probability of 50 percent to obtain this item score or higher

Figure 11.2: Mapping of CT student abilities and item difficulties



For both CIL and CT assessments, the range of item difficulties broadly matched students' abilities found in the student population. However, it is important to acknowledge that the match between test item difficulty and student ability varied considerably across countries depending on the distribution of student achievement within each ICILS 2018 country (see Fraillon et al. 2020, p. 75 & 103).

Assessment of item fit

Before reviewing the international scales in detail and more specific item statistics, we evaluated the model fit for individual items. One way to determine goodness of fit is by calculating a mean square statistic (Wright and Masters 1982). Reviewing this residual-based item fit provides an indication of the extent to which each item fits the item response model. However, there are no clear rules for acceptable item fit, and some statisticians recommend that analysts and researchers interpret residual-based statistics with caution (see, for example, Rost and von Davier 1994).

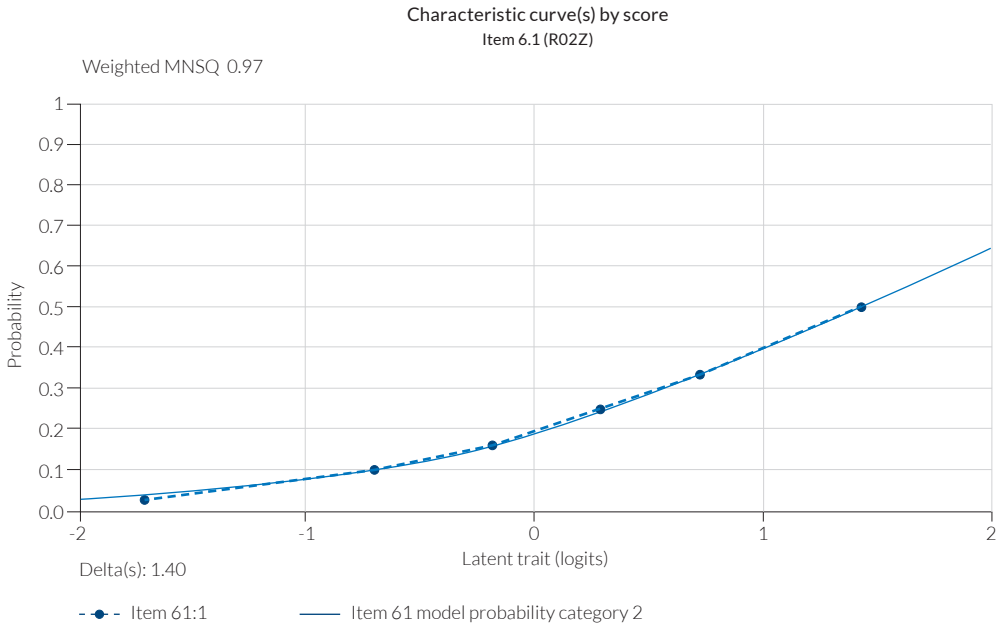
To assess the assumptions of the IRT model, our review of item fit was based on a combination of assessments, such as: mean square statistics, the item-rest correlations, item characteristic curves, percentages of students in each response category, and the average ability of students in each response category. We also reviewed item characteristic curves (ICCs), which provide a graphical representation of the observed success of students on an item in comparison to the probability of success predicted by the model across the range of student abilities for each item, including dichotomous and partial credit items.

While the theory and principles underpinning the evaluation of model fit for items relates to all items, there are slight differences in terms of assessing the psychometric characteristics of items depending on whether they are dichotomous or have more than two categories. In the following section we present examples of reviews of psychometric characteristics for each of these two item types.

Example dichotomous item

Figure 11.3 shows the ICC for item R02Z, which is a dichotomously scored constructed response item. It can be observed that the curve fits the expected model very closely, which is also suggested by the weighted MNSQ of 0.97. The item (location) parameter of 1.40 indicates that this item is moderately difficult, and it was retained for the scaling of CIL items.

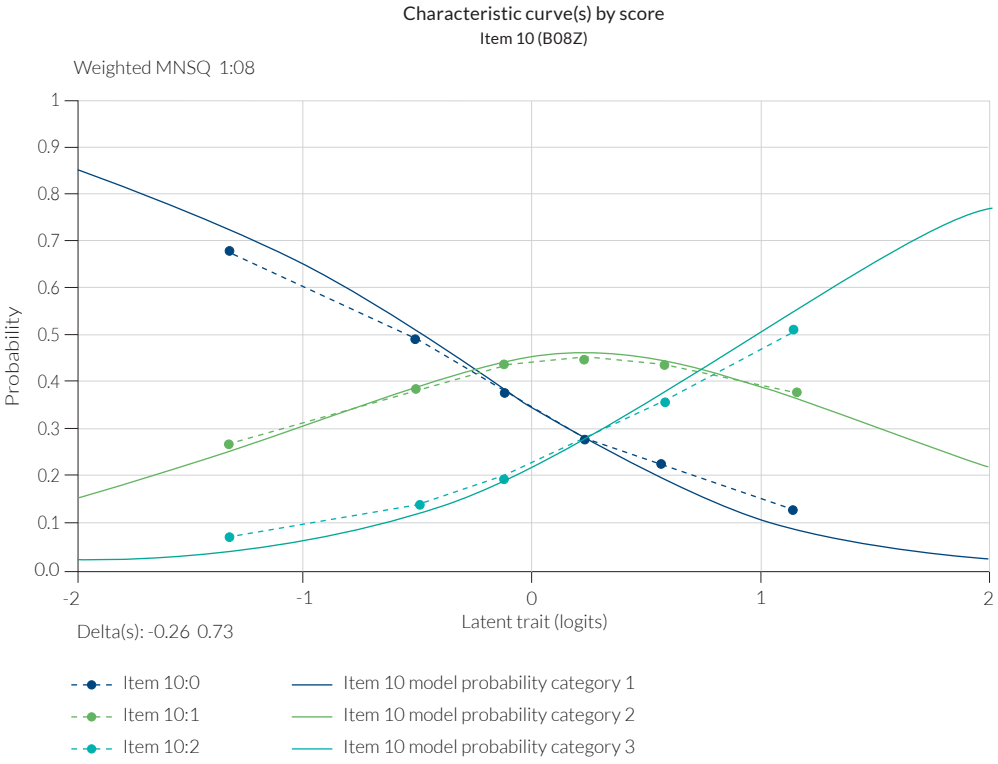
Figure 11.3: Item characteristic curve by score for dichotomous item R02Z



Example partial credit item

The ICC for item B08Z, an item with 0, 1, or 2 score points, showed that, for scores of 0 and 2 score points the curves depicting the observed item responses were not entirely consistent with those predicted by the Rasch partial credit model. Fewer of the students with lower levels of CIL received a score of 0, and more received a score of 2 than predicted by the model. In addition, fewer of the students of higher ability had a score of 2, and more had a score of 0 than predicted. This indicates that the item discriminated not as well as expected. However, the ICC still indicated that students with higher levels of knowledge received higher scores for this item (Figure 11.4). The item was retained for the scaling of CIL items given that its discrimination was judged as still appropriate for the measurement of the latent trait.

Figure 11.4: Item characteristic curve by category for item B08Z



We further analyzed the functioning of the constructed response scoring guides by reviewing the proportion of responses in each score (to confirm that each score was represented in the scoring) and confirming that the mean abilities of students achieving each score on an item (e.g., 0, 1, 2) discernably increased with the increase in scores (i.e., that the mean ability of students achieving a score of 2 on a given item was higher than that of students achieving a score of 1 and that this was in turn higher than that of students receiving a score of 0). This analysis confirmed that all the constructed-response items included in the final set of scaled scored items were of satisfactory psychometric quality.

Item adjudication outcomes

In total, three CIL items (H04A, S05Z, and G05A) were not included in the calibration of items or scaling of students' CIL score. H04A and S05Z had already been excluded from analysis in 2013, but were re-administered in 2018, while G05A was newly developed for ICILS 2018. For these three items, preliminary analysis using 2018 data showed unsatisfactory item statistics and it was decided to exclude these items from further analysis. One further CIL item (S07Z) was excluded from the scaling of CIL items at a later analysis stage due to unsatisfactory scaling properties.

We determined the item-rest correlations, the correlation between the score of one item and the total raw score derived from all other items assigned to each student, of correct responses (or partial credit responses) and the weighted item fit statistics (Table 11.1). Only five CIL items and none of the CT items had an item-rest correlation below 0.2 (which indicates a rather low discrimination). We found unsatisfactory residual-based item fit statistics for only one CIL item and three CT items.

Table 11.1 and 11.2 show the item-rest correlations of correct responses to multiple-choice items or scored partial credit items, and the weighted item fit statistics for CIL and CT items, respectively. Information about item S07Z is still included in this table even though the item was later removed from scaling.

For CIL items, the item-rest correlations ranged from 0.11 to 0.63, while CT items had values ranging from 0.24 to 0.70. Item-rest correlations of 0.20 or lower were usually flagged for further review (six items). Two of the CIL items (G04Z and S06Z) were included in the scaling of CIL items even though we found weighted MNSQ statistics of around 1.20 or higher suggesting somewhat less satisfactory item fit to the model, denoted by less discrimination between high and low performing students than predicted by the model. Similar observations were made regarding the CT items TA05Z and TA07Z that also showed relatively poor fit with weighted MNSQ above 1.20.

Table 11.1: International CIL item-rest correlations and weighted item fit

Item no.	Item name	Item-rest correlation	Weighted fit	Item no.	Item name	Item-rest correlation	Weighted fit
1	B01Z	0.31	1.11	42	S08C	0.56	1.00
2	B02Z	0.38	0.99	43	S08D	0.53	0.99
3	B03Z	0.35	0.98	44	S08E	0.54	0.88
4	B04Z	0.25	1.08	45	S08F	0.55	0.83
5	B05Z	0.36	1.02	46	S08G	0.50	1.02
6	B06Z	0.28	1.03	47	G01Z	0.21	1.10
7	B07A	0.23	1.14	48	G02Z	0.27	1.10
8	B07B	0.30	0.94	49	G03Z	0.11	0.97
9	B07C	0.36	1.06	50	G04Z	0.20	1.17
10	B08Z	0.44	1.12	51	G05B	0.42	0.97
11	B09A	0.58	0.82	52	G06Z	0.25	1.08
12	B09B	0.63	0.89	53	G07Z	0.31	1.13
13	B09C	0.61	0.93	54	G08Z	0.25	1.11
14	B09D	0.49	0.90	55	G09C	0.43	1.02
15	B09E	0.43	1.00	56	G09D	0.44	0.95
16	B09F	0.60	0.83	57	G09E	0.39	0.92
17	B09G	0.55	0.86	58	G09F	0.36	0.87
18	H01Z	0.35	1.04	59	G09G	0.44	0.96
19	H02Z	0.38	1.03	60	R01Z	0.48	0.93
20	H03Z	0.42	0.91	61	R02Z	0.35	1.00
21	H05Z	0.31	1.08	62	R03Z	0.34	1.06
22	H06Z	0.39	0.99	63	R04Z	0.38	1.01
23	H07A	0.55	0.93	64	R05A	0.58	1.17
24	H07B	0.58	0.89	65	R06A	0.34	1.06
25	H07C	0.60	0.94	66	R06B	0.34	1.01
26	H07D	0.49	1.17	67	R07Z	0.41	1.03
27	H07E	0.43	1.15	68	R08Z	0.37	1.02
28	H07F	0.50	0.92	69	R09Z	0.36	1.04
29	H07G	0.40	0.99	70	R10Z	0.25	1.04
30	H07H	0.57	0.97	71	R11A	0.53	1.04
31	H07I	0.48	0.88	72	R11B	0.51	0.86
32	H07J	0.29	0.90	73	R11C	0.53	0.84
33	S01Z	0.23	1.07	74	R11D	0.46	0.93
34	S02Z	0.23	1.12	75	R12A	0.19	1.09
35	S03Z	0.21	1.16	76	R12B	0.37	1.00
36	S04A	0.29	1.04	77	R12C	0.44	0.94
37	S04B	0.34	1.02	78	R12D	0.37	1.01
38	S06Z	0.15	1.21	79	R12G	0.26	1.03
39	S07Z	0.15	1.16	80	R12H	0.58	0.98
40	S08A	0.52	0.87	81	R12I	0.48	0.94
41	S08B	0.59	0.82	82	R12J	0.30	0.87

Table 11.2: International CT item-rest correlations and weighted item fit

Item no.	Item name	Item-rest correlation	Weighted fit
1	TA01Z	0.46	1.12
2	TA02Z	0.41	1.13
3	TA03Z	0.44	1.18
4	TA04Z	0.24	1.09
5	TA05Z	0.37	1.21
6	TA06Z	0.54	0.98
7	TA07Z	0.42	1.27
8	TA08Z	0.34	1.12
9	TF01E	0.54	0.96
10	TF02L	0.48	0.86
11	TF03TEC	0.65	0.84
12	TF04TEC	0.68	0.80
13	TF05TEC	0.70	0.84
14	TF06TEC	0.69	0.80
15	TF07TEC	0.67	0.83
16	TF08TEC	0.54	0.91
17	TF09T	0.44	0.88

Dimensionality and local dependence

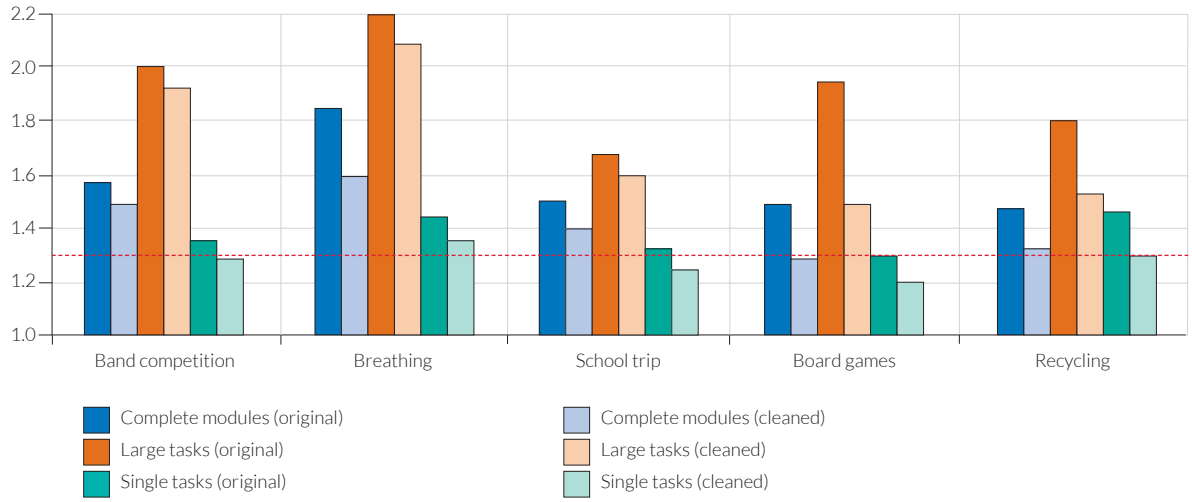
We reviewed test dimensionality for CIL with multidimensional IRT modeling in ICILS 2013 using the ACER ConQuest software package (Adams et al. 2015). The results suggested that students' CIL was best described using a single scale given that latent correlations between potential sub-dimensions were very high (see Gebhardt and Schulz 2015). Analyses of the CT items provided a similar result and only the overall CT scale was reported in ICILS 2018.

An analysis of local dependence was conducted by estimating fit statistics for user-defined combinations of items (Adams and Wu 2009), similar to the fit statistics that are estimated for each item parameter estimate. The expected fit statistics for groups of items was 1, suggesting absence of local dependence between items. A fit statistic of 1.3 or higher was flagged as indicative of local dependence.

The results from the main survey are presented in Figure 11.5. Fit statistics for complete modules (before any action to resolve local dependence was taken) were between 1.5 and 1.9. In all modules, local dependence was highest for items nested within the large tasks. Although less than for items nested within large tasks, fit values above 1.3 indicated that a certain level of local dependence remained for single tasks in four out of five modules. This finding was not unexpected given the structure of the assessment with many items corresponding to common themes or tasks, and similarly high levels of local dependence had already been identified for the CIL assessment in ICILS 2013 (Gebhardt and Schulz 2015). Similar observations have also been made with regard to the IEA Progress in International Reading Literacy Study (PIRLS), where sets of test items are related to common reading passages (Quittre and Monseur 2010).

Analyses to review local dependence were also undertaken between pairs of CIL items within the single tasks or the large task of each module. The purpose of this review was to identify pairs of items with high local dependence in order to collapse those items into one item (if justifiable according to the item content) or, where necessary, to remove one of the two items from the scale. Eleven such pairs were found (with a fit mean square statistic > 1.3). Table 11.3 lists these item pairs and the action taken to resolve the local dependence. Comparison of the fit statistics before and after

Figure 11.5: Local dependence within CIL modules



removing items with unsatisfactory psychometric characteristics showed improvement but local dependence was still evident within four out of five complete modules. When considering only large task items, local dependence was still evident within all five modules, but only within two of five modules when considering only single task items.

Table 11.3: CIL item pairs showing local dependence

Module	Task type	Item 1	Item 2	MNSQ	T value	Action taken
G	large	G09C	G09D	1.36	29	G09C and G09G combined
G	large	G09C	G09G	1.56	42	G09C and G09G combined
H	large	H07A	H07D	1.42	33	H07D removed
H	large	H07C	H07D	1.40	31	H07D removed, H07B, and H07C combined
H	large	H07D	H07F	1.33	25	H07D removed
H	large	H07D	H07G	1.33	25	H07D removed
H	large	H07H	H07I	1.42	32	H07H and H07I combined
R	single	R11A	R11B	1.36	24	R11A , R11B, R11C, and R11D combined
R	single	R11A	R11C	1.43	28	R11A , R11B, R11C, and R11D combined
R	large	R12B	R12C	1.32	28	R12B and R12C combined
S	single	S04A	S04B	1.37	31	S04A and S04B combined

Assessment of scorer reliabilities

The scoring of constructed-response items in the ICILS cognitive test was guided by the scoring guides that were developed for new ICILS 2018 items, then refined following the experiences in the international field trial, or adapted from ICILS 2013 for those items that were included to measure changes over time. Within countries, subsamples of about 20 percent of student responses to each task were scored twice following a controlled, random allocation to different scorers. The assignment of item responses to scorers was implemented and controlled as part of the online scoring systems (see Chapter 3). This double-scoring procedure allowed for the assessment of scorer reliabilities.

Table 11.4 shows the percentages of scorer agreement for CIL and CT items, based on 13 countries and benchmarking entities for CIL and eight countries and benchmarking entities for CT items. Percentage of agreement between scores on double-scored items ranged from 40 to 100 percent across countries.

Table 11.4: Percentages of scorer agreement for constructed-response CIL and CT test items

Country	B01Z	B02Z	B09A	B09B	B09C	B09D	B09E	B09F	B09G	H05Z	H06Z	H07A	H07B	H07C	H07D
Chile	92	94	99	96	93	94	99	98	97	96	98	97	92	93	95
Denmark	83	93	90	88	80	83	97	82	74	97	99	96	80	71	87
Finland	97	99	95	95	95	91	98	91	92	99	98	98	93	92	93
France	87	88	79	86	76	91	92	88	80	94	95	91	51	55	54
Germany	84	94	93	92	78	86	98	84	73	89	96	95	73	69	63
Italy	92	94	97	93	95	88	97	94	92	98	99	99	93	90	89
Kazakhstan	93	96	98	89	94	96	97	98	97	96	98	94	93	92	92
Korea, Republic of	99	99	97	97	97	98	100	98	98	97	99	96	97	98	99
Luxembourg	90	93	95	92	88	97	97	91	87	93	98	100	93	90	92
Moscow (Russian Federation)	93	90	96	94	91	93	97	88	91	93	97	98	90	89	91
Portugal	100	100	96	91	91	100	96	94	87	100	100	96	45	73	86
United States	90	97	97	94	92	87	98	93	88	93	98	94	77	74	75
Uruguay	78	92	92	89	81	79	82	89	86	98	98	90	53	55	56
International ICILS 2018	**91**	**95**	**95**	**92**	**89**	**92**	**96**	**93**	**90**	**96**	**98**	**96**	**81**	**82**	**84**

Country	H07E	H07F	H07G	H07H	H07I	H07J	S08A	S08B	S08C	S08D	S08E	S08F	S08G	G01Z	G03Z
Chile	93	94	95	93	92	91	96	98	99	96	95	92	93	94	92
Denmark	78	86	85	85	84	82	91	98	94	90	94	78	98	94	79
Finland	94	94	94	95	95	94	97	99	99	97	99	93	89	97	91
France	62	53	71	75	82	75	93	97	87	89	91	63	82	91	58
Germany	86	58	84	89	94	81	89	99	96	98	94	71	95	93	63
Italy	92	94	95	98	98	95	98	99	98	98	98	94	95	97	89
Kazakhstan	96	91	90	93	96	98	95	99	97	95	97	94	97	96	89
Korea, Republic of	100	98	98	97	99	97	99	100	99	99	100	98	100	99	97
Luxembourg	100	89	91	93	94	90	96	99	96	97	97	90	99	92	86
Moscow (Russian Federation)	98	94	93	98	99	97	98	99	98	98	99	93	98	92	96
Portugal	73	64	63	75	85	95	99	99	99	97	96	94	100	100	100
United States	62	84	82	86	89	84	95	98	98	97	97	87	89	98	87
Uruguay	77	58	66	62	65	92	89	99	96	92	78	73	74	90	59
International ICILS 2018	**88**	**82**	**86**	**88**	**91**	**90**	**95**	**98**	**97**	**96**	**95**	**87**	**94**	**95**	**83**

Table 11.4: Percentages of scorer agreement for constructed-response CIL and CT test items (contd.)

Country	G06Z	G08Z	G09C	G09D	G09E	G09F	G09G	R02Z	R06A	R06B	R08Z	R10Z	R11A	R11B	R11C
Chile	94	91	93	93	94	96	92	90	94	98	93	92	100	93	99
Denmark	86	83	71	91	93	98	85	84	81	95	85	86	99	94	98
Finland	95	94	91	95	98	98	96	96	94	99	95	97	99	94	99
France	82	72	61	89	91	84	61	78	67	93	77	79	98	79	97
Germany	86	75	72	90	98	98	83	80	84	92	83	81	100	90	N/A
Italy	94	92	93	96	98	99	95	91	91	96	92	90	97	92	97
Kazakhstan	93	94	87	93	97	97	88	92	91	97	92	94	99	94	99
Korea, Republic of	98	99	98	99	96	97	97	98	98	96	96	96	99	97	99
Luxembourg	92	89	88	93	94	95	92	93	91	94	89	88	98	90	95
Moscow (Russian Federation)	96	89	90	92	97	91	95	96	94	97	96	96	99	98	99
Portugal	100	100	89	94	96	96	94	100	96	96	100	100	97	98	97
United States	88	86	85	97	98	97	94	88	85	96	89	93	100	98	100
Uruguay	85	68	63	52	85	84	54	82	40	92	67	64	97	93	99
International ICILS 2018	**92**	**88**	**84**	**91**	**95**	**95**	**87**	**90**	**87**	**96**	**89**	**89**	**99**	**92**	**92**

Country	R11D	R12A	R12B	R12C	R12D	R12G	R12H	R12I	R12J	TA07Z	TA08Z
Chile	96	99	93	91	89	97	91	93	96	N/A	N/A
Denmark	99	95	71	67	88	96	77	89	96	81	74
Finland	99	98	94	90	96	98	97	94	99	95	91
France	76	95	75	68	65	96	84	78	87	79	71
Germany	97	64	73	87	81	94	87	83	94	77	75
Italy	97	99	95	88	94	100	97	98	99	N/A	N/A
Kazakhstan	99	92	92	92	91	93	91	92	91	N/A	N/A
Korea, Republic of	100	99	98	98	98	98	96	97	98	97	96
Luxembourg	96	99	89	86	93	99	96	95	96	88	86
Moscow (Russian Federation)	98	96	92	89	97	97	96	95	95	N/A	N/A
Portugal	100	97	81	83	95	96	93	92	97	100	100
United States	99	97	82	95	89	98	94	93	98	88	87
Uruguay	89	96	82	71	87	97	89	91	85	N/A	N/A
International ICILS 2018	**95**	**94**	**86**	**85**	**89**	**97**	**91**	**92**	**95**	**88**	**85**

As has been the practice in other IEA studies, only items scored with a minimum of 70 percent scorer agreement were included in the international database. While scorer agreement was above 70 percent for all constructed-response items for the pooled ICILS 2018 dataset, we also reviewed and adjudicated scorer agreement at the national level. There were 37 cases where the scoring of an open-ended response item had an agreement below 70 percent. While in eight out of 13 countries scorer agreement was above 70 percent for all constructed-response items, in some countries this criterion had not been met for several of these items. The scores for the corresponding items were excluded from the scaling of the corresponding national data when drawing plausible values but are included in the public database (see Appendix C).

Differential item functioning by gender

The analysis included an exploration of the quality of the items by assessing differential item functioning (DIF) by gender. DIF occurs when groups of students with the same degree of ability differ in their probabilities of responding correctly to an item. For example, if boys with the same degree of ability as girls have a higher probability of correctly answering an item than girls, the item shows DIF with regard to gender. This suggests a violation of the model's assumptions, which assumes that the probability is exclusively a function of ability and not of any other characteristics of the respondents.

It is possible to derive estimates of gender DIF by including interaction terms in the item response model. To achieve this, gender DIF was modeled for dichotomous items as:

$$P_i(\theta_n) = \frac{\exp(\theta_n - (\delta_i - \eta_g + \lambda_{ig}))}{1+\exp(\theta_n - (\delta_i - \eta_g + \lambda_{ig}))}$$

Here, θ_n is the estimated ability of person n and δ_i is the estimated location of item i, an additional parameter for gender effects λ_{ig}. However, to obtain proper estimates, we also needed to include the overall gender effect (η_g) in the model.[3] Both item-by-gender interaction estimates (λ_{ig}) and overall gender effects (η_g) were constrained to have a sum of 0.

Gender DIF estimates for a partial credit model for items with more than two categories (here, constructed items) could similarly be modeled as:

$$P_{X_i}(\theta_n) = \frac{\exp\sum_{j=0}^{x}(\theta_n - (\delta_i - \eta_g + \lambda_{ig} + \tau_{ij}))}{\sum_{h=0}^{m_i}\exp\sum_{j=0}^{h}(\theta_n - (\delta_i - \eta_g + \lambda_{ig} + \tau_{ij}))} \quad x_i = 0,1,\ldots,m_i$$

Here, θ_n denotes the person's ability, δ_i gives the item location parameter on the latent continuum, τ_{ij} is the step parameter, λ_{ig} is the item-by-gender interaction effect, and η_g is the overall gender effect.

Table 11.5 and 11.6 show the gender DIF estimates for CIL and CT items. Estimates above an absolute value of 0.3 logits were flagged as indicating substantial amount of DIF. There were no items that showed indications of DIF that would have suggested a removal from scaling for either scale.

3 The minus sign ensures that higher values of the gender effect parameters indicate higher levels of item endorsement in the gender group with higher value (here, females).

Table 11.5: Gender DIF estimates for CIL test items

Item no.	Item name	Gender DIF estimate	Item no.	Item name	Gender DIF estimate
1	B01Z	-0.04	42	S08C	0.02
2	B02Z	-0.15	43	S08D	-0.10
3	B03Z	-0.11	44	S08E	0.08
4	B04Z	-0.03	45	S08F	0.15
5	B05Z	0.04	46	S08G	0.10
6	B06Z	-0.05	47	G01Z	-0.12
7	B07A	-0.08	48	G02Z	-0.19
8	B07B	0.04	49	G03Z	0.01
9	B07C	-0.04	50	G04Z	-0.16
10	B08Z	-0.10	51	G05B	-0.06
11	B09A	0.15	52	G06Z	-0.02
12	B09B	0.08	53	G07Z	0.09
13	B09C	0.01	54	G08Z	0.09
14	B09D	0.20	55	G09C	-0.04
15	B09E	0.13	56	G09D	-0.01
16	B09F	0.07	57	G09E	0.04
17	B09G	0.13	58	G09F	-0.09
18	H01Z	-0.14	59	G09G	-0.12
19	H02Z	-0.03	60	R01Z	-0.20
20	H03Z	-0.08	61	R02Z	-0.05
21	H05Z	-0.07	62	R03Z	0.01
22	H06Z	0.02	63	R04Z	-0.12
23	H07A	0.09	64	R05A	-0.11
24	H07B	0.06	65	R06A	-0.03
25	H07C	0.04	66	R06B	0.00
26	H07D	-0.03	67	R07Z	0.14
27	H07E	0.02	68	R08Z	0.06
28	H07F	0.00	69	R09Z	0.00
29	H07G	-0.03	70	R10Z	0.13
30	H07H	0.04	71	R11A	0.06
31	H07I	-0.01	72	R11B	0.04
32	H07J	0.03	73	R11C	0.08
33	S01Z	0.00	74	R11D	0.03
34	S02Z	0.00	75	R12A	-0.05
35	S03Z	-0.01	76	R12B	-0.04
36	S04A	-0.04	77	R12C	-0.05
37	S04B	-0.05	78	R12D	0.02
38	S06Z	-0.10	79	R12G	0.00
39	S07Z	-0.17	80	R12H	0.01
40	S08A	0.12	81	R12I	-0.04
41	S08B	0.14	82	R12J	-0.08

Table 11.6: Gender DIF estimates for CT test items

Item no.	Item name	Gender DIF estimate
1	TA01Z	-0.02
2	TA02Z	0.09
3	TA03Z	0.12
4	TA04Z	-0.04
5	TA05Z	0.00
6	TA06Z	-0.03
7	TA07Z	0.07
8	TA08Z	-0.03
9	TF01E	0.07
10	TF02L	0.00
11	TF03TEC	0.06
12	TF04TEC	0.00
13	TF05TEC	0.01
14	TF06TEC	0.03
15	TF07TEC	-0.09
16	TF08TEC	-0.14
17	TF09T	-0.14

National reports with item statistics

National centers were provided with item statistics (see example for B01Z in Figure 11.6) and requested to review any flags for the respective test items. Flags included cases of unusual correlation (e.g., negative correlations between correct response and overall score) and those showing large differences between national and international item difficulties. They also included open-ended items where the category-total correlations were disordered. In some cases, national centers informed the international study center of translation, scoring, or technical problems that had not been detected during verification. In these cases, we categorized the items as "not administered" in the international database and excluded them from scaling of the corresponding national data.

Working independently from the item reviews by national centers, international study center staff flagged national items that showed poor scaling properties (such as item misfit or large item-by-country interactions) and conducted post-verifications of item translation. In one instance, we identified a national item that needed to be set to "not administered" in the international database and was consequently also excluded from the scaling of the corresponding national data.

Appendix C provides details about items that were excluded from scaling or deleted from the database.

Cross-national measurement equivalence

With any test used to assess student achievement cross-nationally, it is important that the test items function similarly across those countries. Similar to the case of DIF by gender (see above), items show item-by-country interaction when students from different countries but with the same ability vary in their probability of answering these questions correctly. Test items with considerable item-by-country interaction are not suitable for the scaling of cognitive test items in international surveys.

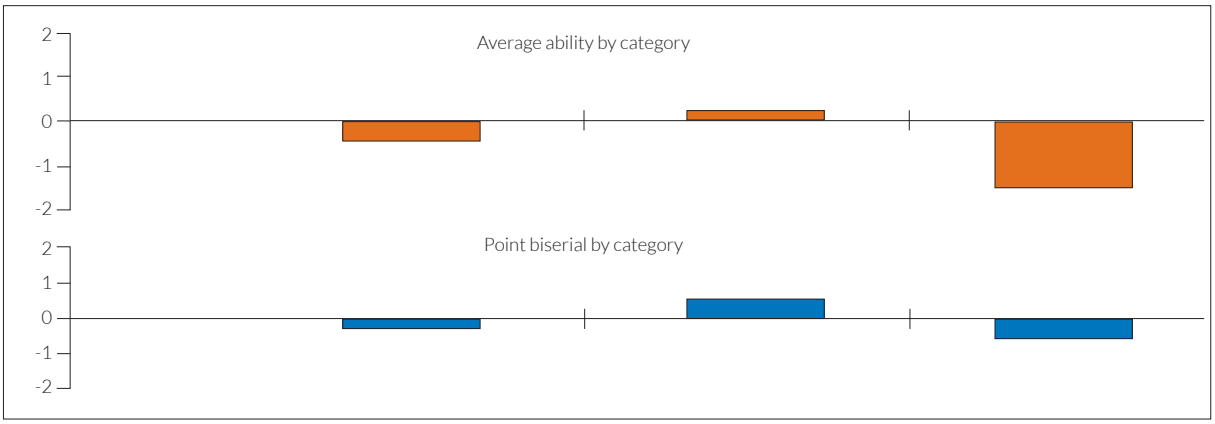
For the main survey analyses of test items, national item parameter calibrations were compared with international item parameters in order to assess the occurrence of item-by-country interaction.

Figure 11.6: Example of item statistics provided to national centres

B01Z: Band competition – 01 (Item Format: constructed response auto-coded)

Number of cases: 1153 Adjusted correlation: 0.18 Item threshold(s): -1.257
Fit (weighted MNSQ): 1.13 Item delta(s) -1.257

Response	0	1	9
Score	0	1	0
Students	223	896	34
% NAT	19.34	77.71	2.95
% INT	33.9	60.35	5.75
Ability average	-0.37	0.061	-1.48
Ability SC	1.084	0.968	0.975
Pt Bis	-0.1	0.19	-0.23



	Fit				Adjusted correlation			
	0.70	1.00	1.30	(value)	0.00	0.40	0.80	(value)
International value		X		1.09		X		0.323
Aggregated statistics								
National value		X		1.13		X		0.184

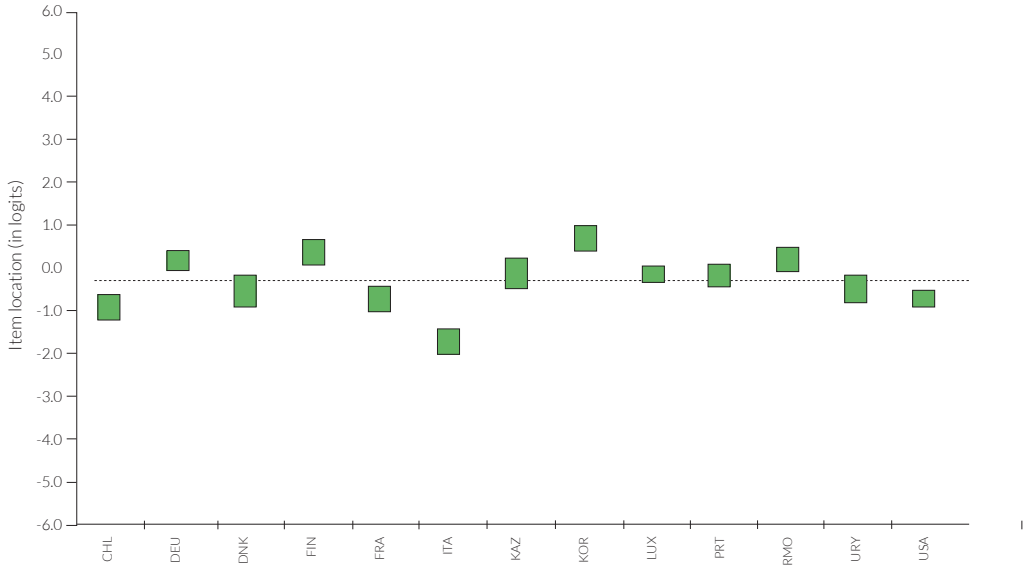
	Delta (item difficulty)				Item-category threshold			
	-2.0	0.0	2.0	(value)	-2.0	0.0	2.0	(value)
International value		X		-0.969		X		-0.969
Aggregated statistics								
National value		X		-1.257		X		-1.257

	Item by country interaction			Adjusted correlation			Fit		
	No of countries	Easier than expected	Harder than expected	Non-key PB is positive	Key PB is negative	Low adjusted correpation	Ability not ordered	Small (high discr.)	Large (low discr.)
B01Z		☐	☐	☐	☐	☑	☐	☐	☐
Countries	13	4	4	0	0	2	0	0	0

Confidence intervals were computed for each national item parameter, basing the computation on the respective standard errors and adjusting them for possible design effects and for multiple comparisons.

As an example, Figure 11.6 shows the item-by-country interaction graph for CIL item H07E. The figure shows clear and considerable variation in the relative difficulty of the item across countries. Similar graphs produced for each test item were used in the test-item adjudication process at the international and national levels, while information about occurrence of cross-national DIF was used to identify items for post-verification checks after completion of the main data collection.

Figure 11.7: Example of item-by-country interaction graph for item H07E



Although the ICILS test items showed generally only limited item-by-country interactions, there were some national item difficulties that deviated quite considerably (more than 1.3 logits) from the international item difficulty. In these cases (see, for example, Italy in Figure 11.7), we omitted the items from scaling. Appendix C includes the complete list of items that were omitted nationally from scaling because of substantial item-by-country interaction.

Evaluating the impact of missing responses

There were two possible types of missing responses in the ICILS test. These were "omitted" items (coded in the database as 9) and "not-administered" items (coded as 8). The omitted-response category was used when a student provided no response at all to an item administered to him or her. Not-administered items were those that, although in the whole item pool, were not in the sets of modules (two out of four) administered to a student. Not-administered items occurred either by design (due to the assignment and rotation of modules) or, in the few cases described earlier in this chapter, due to technical failure, incorrect translations, or scaling properties.

A separate missing category called "not reached" (coded as 7) was created for analysis and subsequent scaling purposes at the post-processing stage. An item was assigned this code if the student concerned did not respond to the item immediately preceding it, and also did not respond to any of the items following this item within the same module (i.e., did not continue on to the end of the test). The extent of occurrence of Code 7 items provided us with information about the eventual appropriateness of the test length as well as the appropriateness of its difficulty, following similar analysis at the field trial stage.

Table 11.7 and 11.8 show the international percentages of omitted and not reached responses.

Table 11.9 shows the average percentages of missing values overall and for each module. On average, each item was omitted by nearly eight percent of the students. The average number of students that did not reach an item was about one percent.

Table 11.7: Percentages of omitted responses and items not reached due to lack of time for CIL items

Item no.	Item name	Omitted	Not reached	Item no.	Item name	Omitted	Not reached
1	B01Z	5.73	0.00	42	S08C	3.33	1.94
2	B02Z	1.32	0.00	43	S08D	3.33	1.94
3	B03Z	25.45	0.02	44	S08E	3.33	1.94
4	B04Z	7.05	0.04	45	S08F	3.33	1.94
5	B05Z	1.35	0.06	46	S08G	3.33	1.94
6	B06Z	7.92	0.10	47	G01Z	2.41	0.00
7	B07A	1.35	0.25	48	G02Z	3.46	0.00
8	B07B	1.36	0.25	49	G03Z	5.64	0.04
9	B07C	1.35	0.25	50	G04Z	1.14	0.06
10	B08Z	1.93	0.33	51	G05B	3.36	0.10
11	B09A	4.11	0.95	52	G06Z	9.43	0.15
12	B09B	4.11	0.95	53	G07Z	0.64	0.30
13	B09C	4.11	0.95	54	G08Z	0.46	0.37
14	B09D	4.11	0.95	55	G09C	25.94	0.64
15	B09E	4.11	0.95	56	G09D	25.94	0.64
16	B09F	4.11	0.95	57	G09E	25.94	0.64
17	B09G	4.11	0.95	58	G09F	25.94	0.64
18	H01Z	2.97	0.00	59	G09G	25.94	0.64
19	H02Z	1.15	0.00	60	R01Z	1.27	0.00
20	H03Z	3.75	0.05	61	R02Z	10.52	0.00
21	H05Z	4.73	0.89	62	R03Z	6.36	0.02
22	H06Z	2.78	1.40	63	R04Z	9.04	0.02
23	H07A	11.82	1.88	64	R05A	3.35	0.04
24	H07B	11.82	1.88	65	R06A	7.02	0.08
25	H07C	11.82	1.88	66	R06B	7.02	0.08
26	H07D	11.82	1.88	67	R07Z	4.82	0.23
27	H07E	11.82	1.88	68	R08Z	11.68	0.30
28	H07F	11.82	1.88	69	R09Z	1.55	0.61
29	H07G	11.82	1.88	70	R10Z	1.19	0.70
30	H07H	11.82	1.88	71	R11A	13.24	0.92
31	H07I	11.82	1.88	72	R11B	13.24	0.92
32	H07J	11.82	1.88	73	R11C	13.88	1.01
33	S01Z	6.68	0.00	74	R11D	13.24	0.92
34	S02Z	1.99	0.00	75	R12A	7.38	5.41
35	S03Z	34.10	0.02	76	R12B	7.38	5.41
36	S04A	1.17	0.05	77	R12C	7.38	5.41
37	S04B	1.17	0.05	78	R12D	7.38	5.41
38	S06Z	1.82	0.40	79	R12G	7.38	5.41
39	S07Z	7.02	0.52	80	R12H	7.38	5.41
40	S08A	3.33	1.94	81	R12I	7.38	5.41
41	S08B	3.33	1.94	82	R12J	7.38	5.41

Table 11.8: Percentages of omitted responses and items not reached due to lack of time for CT items

Item no.	Item name	Omitted	Not reached
1	TA01Z	12.94	0.00
2	TA02Z	10.24	0.00
3	TA03Z	21.15	0.17
4	TA04Z	2.94	0.32
5	TA05Z	3.69	0.36
6	TA06Z	2.81	0.75
7	TA07Z	11.04	2.09
8	TA08Z	5.35	8.19
9	TF01E	1.84	0.00
10	TF02L	3.33	0.00
11	TF03TEC	4.01	0.40
12	TF04TEC	2.96	0.67
13	TF05TEC	3.82	1.21
14	TF06TEC	17.40	2.21
15	TF07TEC	8.33	6.07
16	TF08TEC	30.69	10.14
17	TF09T	18.63	30.43

Table 11.9: Percentages of omitted responses and items not reached due to lack of time overall, by module

	Average percentage	
	Omitted	Not reached
Module		
Band competition	4.9	0.5
Breathing	8.9	1.4
School trip	5.5	1.0
Board games	12.0	0.3
Recycling	7.7	2.1
Grand average	**7.6**	**1.2**

When comparing the proportion of omitted and not reached responses across CIL modules, we observed that the *Board games* module had the highest percentage of omitted responses (12%), while *Band competition* module had the lowest proportion (5%). Average percentages of not-reached responses were close to zero for most modules. *Recycling* module showed somewhat higher percentages of not-reached responses than the other modules.

International item calibration and test reliability

Item parameters for CIL were obtained from a joint data file that included response data from both ICILS 2013 and ICILS 2018. We included the ICILS 2013 data to improve the estimation of link items and for the purpose of equating, using the preferred IEA joint calibration methodology also applied for the equating of TIMSS, PIRLS, and ICCS data (see Foy and Yin 2016, 2017; Gebhardt and Schulz 2018). We included ICILS 2018 data from all 11 countries that met the sampling requirements and ICILS 2013 data from three countries that participated in both 2013 and 2018 and had met sample participation requirements. Denmark participated in both 2013 and 2018 but did not meet sample participation requirements in 2018, and so was not included in the joint data file. Countries were equally weighted within each ICILS cycle for the CIL calibration and all items were included (except for items that were deleted nationally or internationally following the adjudication process).

Table 11.10: Final parameter estimates of the CIL items

Item no.	Item name	Item parameter	Step 1	Item no.	Item name	Item parameter	Step 1	Step 2
1	B01Z	-0.939		42	S08C	0.017	3.145	
2	B02Z	-0.895		43	S08D	0.231	0.025	
3	B03Z	0.168		44	S08E	0.632	0.911	
4	B04Z	0.701		45	S08F	0.931	-0.653	
5	B05Z	-0.016		46	S08G	-0.285	-0.085	
6	B06Z	-2.678		47	G01Z	1.424		
7	B07A	-3.766		48	G02Z	-0.372		
8	B07B	-3.326		49	G03Z	1.742		
9	B07C	-0.905		50	G04Z	-1.045		
10	B08Z	0.130	-0.746	51	G05B	-0.487		
11	B09A	0.197	-1.683	52	G06Z	1.046		
12	B09B	-0.712	0.915	53	G07Z	-1.543		
13	B09C	-0.487	0.859	54	G08Z	-0.270		
14	B09D	-0.448		55	G09C	0.255	-0.256	
15	B09E	0.852		56	G09D	0.180		
16	B09F	0.179		57	G09E	1.490		
17	B09G	0.321		58	G09F	1.787		
18	H01Z	-0.648		59	G09G	-0.052		
19	H02Z	0.612		60	R01Z	-0.454		
20	H03Z	-1.509		61	R02Z	1.033		
21	H05Z	1.456		62	R03Z	-0.382		
22	H06Z	0.529		63	R04Z	0.166		
23	H07A	-1.481		64	R05A	-0.201	0.729	-0.875
24	H07B	0.197	0.108	65	R06A	-0.456		
25	H07C	-0.374	-0.015	66	R06B	0.631		
26	H07D	-0.743	0.245	67	R07Z	0.437		
27	H07E	-0.168	0.079	68	R08Z	0.330		
28	H07F	0.300	0.231	69	R09Z	-0.798		
29	H07G	0.945	0.421	70	R10Z	0.647		
30	H07H	-0.128	0.344	71	R11A	-1.305	2.717	
31	H07I	0.291		72	R11B	-1.283		
32	H07J	1.615		73	R11C	-1.226		
33	S01Z	-1.971		74	R11D	-0.029		
34	S02Z	2.665		75	R12A	2.088		
35	S03Z	0.320		76	R12B	0.502		
36	S04A	1.188		77	R12C	-0.708		
37	S04B	-1.006		78	R12D	0.821		
38	S06Z	0.088		79	R12G	2.044		
39	S07Z	2.613		80	R12H	0.402	0.150	
40	S08A	-0.509		81	R12I	1.325		
41	S08B	-0.027		82	R12J	1.870		

Table 11.11: Final parameter estimates of the CT items

Item no.	Item name	Item parameter	Step 1	Step 2
1	TA01Z	-0.234	0.485	
2	TA02Z	-1.109	0.317	
3	TA03Z	-0.646	1.157	
4	TA04Z	0.733		
5	TA05Z	0.233	-0.047	
6	TA06Z	0.017	0.225	
7	TA07Z	1.105	0.711	
8	TA08Z	0.710	-0.377	
9	TF01E	-1.205	-1.352	-0.030
10	TF02L	-1.587	0.818	
11	TF03TEC	-0.278	-0.466	-1.546
12	TF04TEC	-0.578	0.031	-1.497
13	TF05TEC	-0.132	-0.921	0.454
14	TF06TEC	0.635	-0.757	-0.585
15	TF07TEC	0.544	-0.429	0.603
16	TF08TEC	0.700	2.042	
17	TF09T	1.090	1.663	

Item parameters for CT were obtained from a data file that included response data from all seven countries that met the sampling requirements for ICILS 2018.

Missing student responses that were likely to be due to problems with test length ("not reached items") were excluded from the calibration of item parameters, but included and treated as "incorrect" when scaling the student responses. Items for which technical failures occurred were treated as not administered. Omitted items were treated as incorrect at both stages.

From this, we identified a set of item parameters that we used to scale the ICILS 2018 CIL data (Table 11.10). The final set of CT item parameters is displayed in Table 11.11.

The overall test reliabilities for the two cognitive assessments following the removal of items with non-satisfactory psychometric properties, based on the pooled datasets and obtained from the scaling model, were 0.97 (CIL) and 0.84 (CT) (ACER ConQuest 4.0 estimate).

CIL and CT estimates

The accuracy of measuring the latent ability θ at the individual level can be improved by using a larger number of test items. However, in large-scale surveys such as ICILS, the purpose is to obtain accurate population estimates through use of instruments that also cover a wider range of possible aspects of cognitive abilities.

The use of a matrix-sampling design, where individual students are allocated modules in a systematic way and respond to a set of items obtained from the main pool of items, has become standard in assessments of this type (see Chapter 2). However, reducing test length and administering subsets of items to individual students introduces a considerable degree of uncertainty at the individual level. Aggregated student abilities of this type can lead to bias in population estimates. This problem can be addressed by essentially treating a student's ability estimate as a missing data problem and employing plausible value methodology that uses all available information from student tests and questionnaires to impute an ability estimate, a process that leads to more accurate population as well as sub-group estimates (Mislevy 1991; Mislevy and Sheehan 1987; von Davier et al. 2009).

Using item parameters anchored at their estimated values from the calibration sample makes it possible to randomly draw plausible values from the marginal posterior of the latent distribution for each individual. Estimations are based on the conditional item response model and the population model, which includes the regression on background variables and between-school differences used for conditioning. (For a detailed description see Adams et al. 1997; also Adams 2002.) In order to obtain estimates of students' CIL and CT, we used the ACER ConQuest 4.0 software to draw plausible values.

All available international student questionnaire variables were used for conditioning on background information on students. To take missing responses into account, all missing values in a variable were substituted with either the mode or the mean and corresponding indicators for the occurrence of missing values were added as additional variables. Appendix D 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.

Because of the large number of variables, principal component analyses (PCA) were used to reduce the number of student-level variables as conditioning variables so that these reflected 99 percent of the variance in the original variables. At the student level, only gender and its corresponding missing indicator were included as direct conditioning variables. To account for between-school differences, stratum indicators and the average of (weighted likelihood) ability estimates for all other students in the same school were also introduced as direct conditioning variables.

For ICILS 2013, the CIL scale was originally established and transformed to a metric with a mean of 500 and a standard deviation of 100 for equally weighted ICILS 2013 countries that had met sampling requirements (categories 1 and 2), also excluding benchmarking participants (see Gebhardt and Schulz 2015). This linear transformation was computed by applying the formula:

$$\theta'_{n(CIL)} = 500 + 100\left(\frac{\theta_{n(CIL)} - \mu_{\theta(CIL13)}}{\sigma_{\theta(CIL13)}}\right),$$

where $\theta'_{n(CIL)}$ were the student scores in the international metric, $\theta_{n(CIL)}$ were the original logit scores, $\mu_{\theta(CIL13)}$ was the average of students' CIL logit scores (-0.119) for a pooled dataset with equally weighted national ICILS 2013 samples from countries that had met IEA sample participation requirements, and $\sigma_{\theta(CIL13)}$ was the corresponding standard deviation (1.186). This transformation was applied to each of the five plausible values reflecting CIL derived for ICILS 2013. The equating of CIL scores derived in ICILS 2018 will be described in the next section.

To transform the original CT score in ICILS 2018 to a new reporting metric for this cycle, a similar transformation was applied as for establishing the CT scale metric:

$$\theta'_{n(CT)} = 500 + 100\left(\frac{\theta_{n(CT)} - \mu_{\theta(CT13)}}{\sigma_{n\theta(CT13)}}\right)$$

Here, $\theta'_{n(CT)}$ represents the student CT scores in the international metric, $\theta_{n(CT)}$ the original logit scores, $\mu_{\theta(CT13)}$ the average of students' CT logit scores (-0.1490) for a pooled dataset with seven equally weighted national ICILS 2018 samples from countries that had participated in the international option and met IEA sample participation requirements, and $\sigma_{n\theta(CT13)}$ the corresponding standard deviation (0.9702).

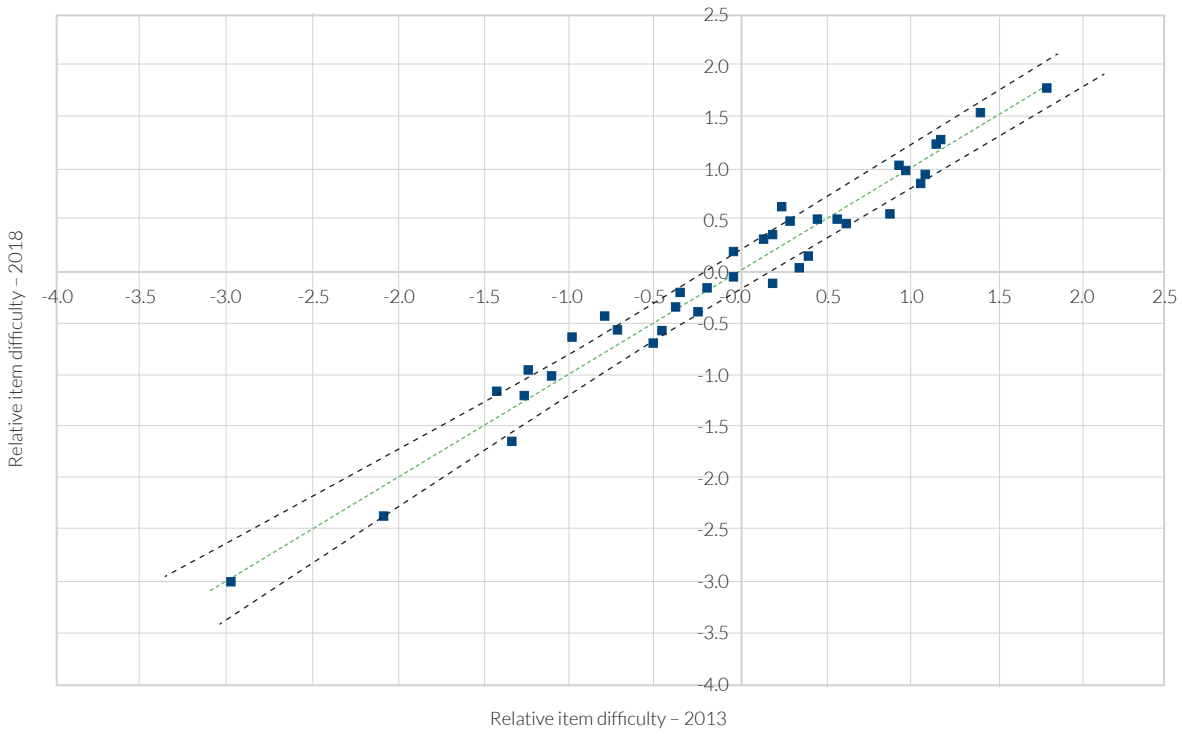
Equating CIL scores from ICILS 2013 and 2018

To achieve the transformation of ICILS 2018 CIL scores to the scale originally established with ICILS 2013 data, it was necessary to equate the new scale scores. As mentioned earlier, all ICILS 2013 item parameters were re-estimated concurrently during the ICILS 2018 joint calibration process.

Before joining the data from the two assessment cycles, we reviewed the relative difficulties of the common items to evaluate the quality of the link. In total there were 46 items common between the two ICILS assessment cycles. To compare relative difficulties of the 46 common items between the two assessments, all ICILS 2018 items were also calibrated separately using the data from 11 countries (excluding two benchmarking entities and the data from the United States as these participants did not meet the sampling requirements). These item difficulties were compared with the item difficulties estimated from the calibration sample in ICILS 2013.

The difference in relative difficulty for each item between the values estimated in 2013 and 2018 calibrations was used to assess the quality of the items as a link. Differences were expected to be zero. A difference of more than half a logit was considered to be large and would result in breaking the link of that item. After a careful consideration of results of various comparisons, a final set of 36 link items was selected. The remaining 10 items common between two cycles were kept in the joint calibration dataset but were treated as separate items.

Figure 11.8: Relative item difficulties for CIL common items in 2013 and 2018



To construct ICILS 2018 scale, the data from 11 countries that met the sampling requirement in 2018 were merged with the data from three trend countries collected in the ICILS 2013 assessment. A concurrent calibration was performed regressing on the cycle to estimate the parameters of all items used over the two assessment cycles. In total the item difficulties were estimated for 111 items. These included all 82 items used in ICILS 2018, 19 items used only in ICILS 2013, and an additional 10 items that were un-linked and thus estimated for 2013 and 2018 separately (no overlap of the data in the combined dataset).

For equating purposes, the new item parameter estimates were used to redraw plausible values for the ICILS 2013 sample, using full conditioning. We only included the three countries that met the sampling requirements in both cycles. Subsequently, we computed the pooled mean and standard deviation of the plausible values on the 2013 scale and on the 2018 scale. Comparing those distributions resulting in the following linear transformation to equate the ICILS 2018 student abilities onto the historical ICILS 2013 scale in logits:

$$\theta_n^E = 1.0258 \times \theta_n^{18} + 0.2127$$

These equated plausible values were subsequently placed on the ICILS 2018 international reporting scale by applying the same transformation as in ICILS 2013:

$$\theta_n' = 500 + 100 \left(\frac{\theta_n^E - (-0.1188)}{1.1859} \right)$$

Because the transformation equating the ICILS 2018 data with the ICILS 2013 data depended on the change in the degree of difficulty of each of the individual link items, the sample of link items chosen influenced the choice of transformation. This meant 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 and Berezner 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.

The following approach has been used to estimate the equating error. 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 $\hat{\delta}_{ij}^{y}$ is the estimated difficulty of item i in unit j for year y, and let:

$$c_{ij} = \hat{\delta}_{ij}^{2018} - \hat{\delta}_{ij}^{2013}$$

The size (number of score points) of unit j is m_j so that:

$$\sum_{j=1}^{K} m_j = L \text{ and } \bar{m} = \frac{1}{K} \sum_{j=1}^{K} m_j$$

Further let:

$$c_{\cdot j} = \frac{1}{m_j} \sum_{i=1}^{m_j} c_{ij}, \text{ and } \bar{c} = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{m_j} c_{ij}$$

and then the link error, taking into account the clustering of items was computed as follows:

$$LinkError_{2018,\,2013} = \sqrt{\frac{\sum_{j=1}^{K} m_j^2 (c_{\cdot j} - \bar{c})^2}{K(K-1)\bar{m}^2}} = \frac{\sum_{j=1}^{K} m_j^2 (c_{\cdot j} - \bar{c})^2}{L^2} \frac{K}{K-1}$$

The development of proficiency levels for CIL

One of the objectives of ICILS was to establish a described CIL scale that would become a reference point for future international assessments in this learning area. Establishing proficiency levels of CIL is an informative way of describing student performance across countries and also sets benchmarks for future surveys.

Students whose results are located within a particular level of proficiency are typically able to demonstrate certain understandings and skills that are associated with that level. These students also typically possess the understandings and skills defined as applying at lower proficiency levels.

When developing proficiency levels, a method was applied that ensured that the notion of "being at a level" could be interpreted consistently and in line with the fact that the achievement scale is a continuum. It was therefore attempted to provide a common understanding about what being at a level meant and to ensure that this meaning was consistent across different proficiency levels. This method took the following three questions into account:

- What is the expected success of a student at a particular level on a test containing items at that level?
- What is the width of the levels in that scale?
- What is the probability that a student in the middle of a level will correctly answer an item of average difficulty for that level?

We adopted the following two parameters for defining proficiency levels to create the properties described below:

- *The response probability (rp) for reporting item parameters:* this was set at *rp* = 0.62 (providing a more appropriate level of "mastery" than *rp* = 0.5).
- *The width of the proficiency levels:* this was set at 0.8 logits (i.e., the original latent trait scores from the IRT scaling model prior to their transformation to the reporting metric).

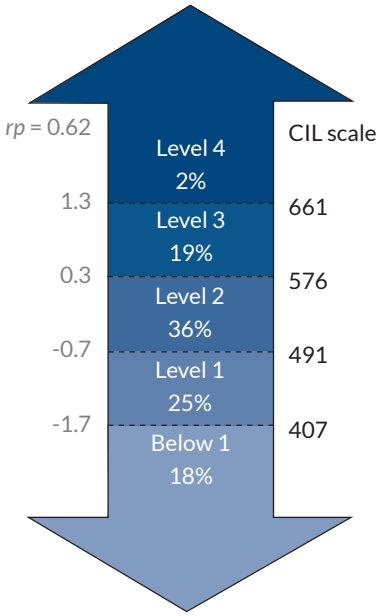
Using these parameters, we were able to infer the following about students' CIL in relation to the proficiency levels:

- Students whose results placed them at the lowest possible point of a proficiency level were likely to correctly answer (on average) slightly over 50 percent of the items on a (hypothetical) test made up of items with locations spread uniformly across the level.
- Students whose results placed them at the lowest possible point of a proficiency level had a 62 percent probability of giving the correct response to an item at the bottom end of the proficiency level.
- Students whose results placed them at the top of the proficiency level had a 78 percent probability of correctly responding to an item at the bottom end of the proficiency level.

The approach that was chosen was essentially an attempt to apply an appropriate choice of mastery by placing item locations at *rp* = 0.62 while simultaneously ensuring that the approach would be understood by the readers of ICILS reports.

The international research team identified and described four proficiency levels that could be used when reporting student performances in CIL from the assessment. Figure 11.9 shows the cut-points for these levels (in logits and final scale scores). The figure also cites the percentage of students at each proficiency level across the participating ICILS countries.

Figure 11.9: CIL proficiency level cut-points and percentage of students at each 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 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 CIL item was obtained as follows:

$$\delta_i' = 500 + 100 \times \left(\frac{(1.0258 \times \delta_i + 0.2127) - \ln\left(\frac{0.62}{0.38}\right)\mu_{\theta(CIL13)}}{\mu_{\theta(CIL13)}} \right)$$

Here, δ_i is the item difficulty in its original metric, $\mu_{\theta(CIL13)}$ is the ICILS 2013 average of students' CIL logit scores (-0.119) and $\mu_{\theta(CIL13)}$ is its corresponding standard deviation (1.186) that were used to standardize the plausible values. As the CIL item difficulty parameters (δ_i) were calibrated based on the combined set of old and new items, the same transformation as for student CIL scores had to be applied before transforming them to the CIL reporting metric at *rp* = 0.62.

References

Adams, R. (2002). Scaling PISA cognitive data. In R. Adams, & M. Wu (Eds.), *Technical report for the OECD Programme for International Student Assessment* (pp. 99–108). Paris, France: OECD Publications.

Adams, R. J., & Wu, M. L. (2009). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalized item response models. *Journal of Applied Measurement*, *10*(4), 1–16.

Adams, R., Wu, M., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematical and science scales. In M. O. Martin, & D. L. Kelly (Eds.), *TIMSS technical report: Vol. II. Implementation and analysis: Primary and middle school years*. Chestnut Hill, MA: Boston College.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software [computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research (ACER).

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* (pp. 13.1–13.62). http://timss.bc.edu/publications/timss/2015-methods/chapter-13.html

Foy, P., & Yin, Y. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 12.1–12.38). https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-12.html

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030193881

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*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030387808

Gebhardt, E., & Schulz, W. (2015). Scaling procedures for ICILS test items. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *International Computer and Information Literacy Study 2013 technical report* (pp. 155–176). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Gebhardt, E., & Schulz, W. (2018). Scaling procedures for ICCS 2016 test items. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 117–137). Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

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). New York/Berlin/Heidelberg: Springer.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *The NAEP 1983–1984 technical report* (Report No. 15-TR-20, pp. 293–360). Princeton, NJ: 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.

Quittre, V., & Monseur, C. (2010). *Exploring local item dependency for items clustered around common reading passage in PIRLS data*. Paper presented at the fourth IEA International Research Conference, Gothenburg, Sweden, 1-3 July. Retrieved from https://www.iea.nl/sites/default/files/2019-04/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement*, *18*(2), 171–182.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *vol. 2*, 9–36.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: Mesa Press.

CHAPTER 12:

Scaling procedures for ICILS 2018 questionnaire items

Wolfram Schulz and Tim Friedman

Introduction

This chapter describes the procedures for the scaling of the ICILS 2018 questionnaire data (collected from students, teachers, ICT coordinators, and school principals) and the indices that were derived from these data.

When describing the ICILS 2018 questionnaire indices, we distinguish the following two general types of indices:

- *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;
- *Scale indices* that were derived through the scaling of items, this was typically achieved by using item response modeling of items with two or more categories.

The chapter is divided into three main sections. The first section lists the simple indices that were derived from ICILS 2018 data and describes how they were created. The second section outlines the procedures used for the scaling of questionnaire data in ICILS 2018 followed by the third section which lists the scaled indices with statistical information on the factor structure of related item sets, scale reliabilities, and parameters used for the item response theory (IRT) scaling.

Results from an analysis of cross-country validity of item dimensionality and constructs were already part of ICILS 2018 field trial analyses. The international study center at ACER conducted reviews of the extent to which measurement models were equivalent across participating countries for draft item material. To conduct this review we made use of exploratory as well as confirmatory factor analysis and item response modeling to examine cross-national measurement equivalence before the final selection of main survey questionnaire items (see examples of this type of analysis in Schulz 2009, 2017).

Simple indices

Student questionnaire

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 the student questionnaire (Question 1) was used to derive age, except for students where this information was missing. In these cases information from student tracking forms (see Chapter 10 for more details) 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., "2018" or "2005"), 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.

In Question 2, students were asked about their gender. These were recorded as the *sex of student* (S_SEX), a girl (1) or a boy (0). For students with omitted data for this question, the gender from the tracking form was included.

Question 3 asked students about their expected highest level of educational attainment. The responses were classified using the International Standard Classification of Education (ISCED) framework (UNESCO 2012) The corresponding index *students' expected education* (S_ISCED) had the following categories:

(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).

The ICILS student questionnaire collected information on the *country of birth* of the students and their parents (Question 4). 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 of *immigrant background* (S_IMMIG) was created using these three indicator variables and had three categories:

(0) Native students (students who had at least one parent born in the country of assessment);

(1) First-generation students (students who were born in the country of assessment and whose parent(s) were born in another country); and

(2) Non-native students (students who were born outside the country of assessment and whose parent(s) were born in another country).

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 from immigrant families (coded 1) and students from non-immigrant families (coded 0) called S_IMMBGR.

Question 5 of the ICILS student questionnaire asked students if the language spoken at home most of the time was the language of assessment or another language.[1] We used this information to derive an index on *home language* (S_TLANG), in which responses were grouped into two categories:

(1) The language spoken at home most of the time was the language of assessment; and

(0) The language spoken at home most of the time differed from the language of assessment.

Occupational data for students' parents[2] were obtained by firstly asking students whether their parents are in paid work or not (Questions 6 and 10), and secondly asking students to provide details about their parents' jobs, using a pair of open-ended questions (Questions 7/8 and 11/12). The paid work status of the students' parents was classified into S_P1WORK (parent 1) and S_P2WORK (parent 2) (1 indicating that the parent is in paid work and 0 indicating that the parent is not in paid work). The open-ended 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

1 Most countries collected more detailed information on language use. This information is not included in the international database.

2 Students could complete parental occupation questions for up to two parents.

and S_P2ISCO for the students' parents. We then mapped these codes to the international socioeconomic index of occupational status (ISEI) (Ganzeboom et al. 1992). The three indices that we obtained from these scores were occupational status of the first parent (S_P1ISEI), occupational status of the second parent (S_P2ISEI), and the highest occupational status of both parents (S_HISEI), with the latter corresponding to the higher ISEI score of either parent or to the only available parent's ISEI score. For all three indices, higher scores indicate higher levels of occupational status.

Questions 9 and 13 asked about the *highest parental education* attainment for the students' parents and provided the data for measuring another important family background variable. The core difficulties with this variable relate to international comparability (education systems differ widely across countries and over time within countries) and response validity (students are often unable to accurately report their parents' levels of education). 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 (S_P1ISCED) and second parent (S_P2ISCED). The index for *highest educational level of parental education* (S_HISCED) corresponds to the higher ISCED level of either parent.

Question 14 of the ICILS student questionnaire asked students how many books they had in their homes. Responses to this question formed the basis for an index of students' *home literacy resources* (S_HOMLIT) with 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.

Question 15B of the ICILS student questionnaire was an international option question where students were asked if they have an *internet connection at home*. The index (S_INTNET) was recorded as Yes (1) or No (0).

In Question 16, students were asked their number of years of experience in using different types of ICT devices. The three types of devices included:

- Desktop or laptop computers;
- Tablet devices or e-readers (e.g., iPad, Tablet PC, Kindle); and
- Smartphones except for using text and calling.

Responses to these items were coded into three respective indices: *computer experience in years* (S_EXCOMP), *tablet experience in years* (S_EXTAB), and *smartphone experience in years* (S_EXSMART) with the following response options available:

(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.

The last question of the student questionnaire, Question 30, asked students whether they studied a computing subject (computing, computer science, information technology, informatics or similar) during the current school year. Responses were used to derive an index of *ICT studies in current school year* (S_ICTSTUD) which was recorded as Yes (1) or No (0).

Teacher questionnaire

The *sex of teacher* (T_SEX) was computed from the data captured from Question 1 of the teacher questionnaire. Female teaches were coded as 1, whereas male teachers were coded as 0.

Teacher age (T_AGE) consisted of the midpoint of the age ranges given in Question 2 of the teacher questionnaire. We assigned "less than 25" a value of 23 and coded "60 or over" as 63.

Question 4 asked teachers to indicate the number of schools in which they teach the target grade population for the current school year. Data captured from this question were used to calculate the *allocation of teacher's staff time to sampled school* (T_WGT) and was coded as:

(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.

Question 5 of the teacher questionnaire asked teachers to indicate how long they have been using computers for teaching purposes. They were asked to distinguish between *ICT experience with ICT use during lessons* (T_EXLES) and *ICT experience with ICT use for preparing lessons* (T_EXPREP). These indices were coded as:

(0) Never;

(1) Less than two years;

(2) Between two and five years; and

(3) More than five years.

School questionnaires

The first question of the principal questionnaire asked whether respondents were male or female. This was used to form the index *sex of principal* (P_SEX) where female principals were coded as 1, and male principals coded as 0.

The ICILS school principal questionnaire asked in Question 3 about the number of girls and boys in the entire school (IP2G03A, IP2G03B) and in Question 4 also about the enrolled girls and boys at the target grade (IP2G04A, IP2G04B). The numbers given for each gender group were summed to form an index of the *number of students in the entire school* (P_NUMSTD) and of the *number of students in the target grade* (P_NUMTAR).[3]

Question 5 also asked principals to report the lowest (youngest) (IP2G05A) grade and the highest (oldest) (IP2G05B) grade taught in their school. The difference between these two grades was calculated as the *number of grades in school* (P_NGRADE).

3 These indices will be included in the ICILS 2018 Restricted Use Data file, while the ICILS 2018 Public Use Data file will include these variables in a categorized form as P_NUMSTD_CAT (1 = 1-300, 2 = 301-600, 3 = 601-900, 4 = more than 900) and P_NUMTAR_CAT (1 = 1-100, 2 = 101-200, 3 = more than 200).

Question 6 collected the information on the *number of teachers* (P_NUMTCH) in a school. The index was calculated by summing the total number of full time teachers (IP2G06A) with the total number of part-time teachers weighted at 50 percent (0.5 x IP2G06B).[4] The *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.

Question 8A collected information about whether the school was a public school or private school. This was used to form a *private school indicator* (P_PRIV) where public schools are coded as 0 and private schools coded as 1.[5]

Question 8B 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). We used the responses to compute an indicator of *school composition by student background* (P_COMP) where a value of 1 was assigned to "schools with more affluent than disadvantaged students," 2 to "schools with neither more affluent nor more disadvantaged students," and 3 to "schools with more disadvantaged than affluent students."

Question 3 of the ICT coordinator questionnaire asked respondents to indicate the *ICT experience in years in the school* (C_EXP). These were recoded 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.

Question 7 of the ICT coordinator questionnaire collected data on the number of desktop computers, the number of laptop/notebooks, and the number of tablet devices in the school. The *sum of ICT devices* (C_ICTDEV) was derived by adding these numbers. 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). The question also asked respondents to indicate the number of school-provided smart boards or interactive white boards 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 (II2G07C).

4 This index will be included in the ICILS 2018 Restricted Use Data, while the ICILS 2018 Public Use Data will include this variable in a categorized form as P_NUMTCH_CAT (1 = 1-25, 2 = 26-50, 3 = 51-75, 4 = more than 75).

5 This index will only be included in the ICILS 2018 Restricted Use Data.

Scaling procedures

Review of item statistics

As part of the scaling analyses of questionnaire data, we reviewed reliabilities both overall and for national samples using Cronbach's alpha coefficient as an estimate of the internal consistency of each scale (Cronbach 1951). When reviewing reliabilities we regarded values above 0.7 as indicating satisfactory internal consistency and those above 0.8 as showing high degrees of reliability (see, for example, Nunnally and Bernstein 1994, pp. 264–265). Apart from scale reliabilities, this analysis stage also considered the percentages of missing responses (which tended to be very low in most cases) as well as the correlations between individual items and the scale scores based on all other items in a scale (adjusted item-total correlations).

Confirmatory factor analysis

Structural equation modeling (SEM) (Kaplan 2009) provides a tool for modeling and confirming theoretically expected dimensions measured with sets of student, teacher, or school questionnaire data. At the field trial stage, it can also be used to re-specify originally expected dimensional structures.[6] When using confirmatory factor analysis, researchers acknowledge the need to employ a theoretical model of item dimensionality that can be tested via the collected data. Within the SEM framework, latent variables link to observable variables via measurement equations. An observed variable *x* is thus modeled as:

$$x = \Lambda_x \xi + \delta$$

where Λ_x is a q x k matrix of factor loadings, ξ denotes the latent variable(s), and δ is a q x 1 vector of unique error variables. The expected covariance matrix is fitted according to the theoretical factor structure.

When conducting the confirmatory factor analyses for ICILS 2018 questionnaire data, selected model-fit indices provided measures of the extent to which a particular model with an assumed a-priori structure had a good fit with regard to the observed data. For the ICILS 2018 analysis, the assessment of model fit was primarily conducted through reviews of the *root-mean square error of approximation* (RMSEA), the *comparative fit index* (CFI), and the *non-normed fit index* (NNFI), all of which are less affected than other indices by sample size and model complexity (see Bollen and Long 1993).

We interpreted RMSEA values indicating model fit as unacceptable with values over 0.10, as marginally satisfactory with values between 0.08 and 0.10, as satisfactory between 0.05 and 0.08, and as a close fit with values lower than 0.05 (see MacCallum et al. 1996). As additional fit indices, CFI and NNFI are bound between 0 and 1. Values below 0.90 indicate a non-satisfactory model fit whereas values greater than 0.95 were interpreted as suggesting a close model fit (see Bentler and Bonnet 1980; Hu and Bentler 1999).

In addition to these fit indices, reviews of standardized factor loadings and the corresponding residual item variances provide further evidence of model fit for questionnaire data. Standardized factor loadings λ' can be interpreted in the same way as standardized regression coefficients by assuming that indicator variables are regressed on an underlying latent factor. The loadings reflect the extent to which each indicator measures the underlying construct. Squared standardized factor loadings indicate how much variance in an indicator variable can be explained by the latent factor and are related to the (standardized) residual variance estimate δ' (reflecting the estimated proportion of unexplained variance) as:

$$\delta' = (1-\lambda'^2)$$

6 In the initial stages of field trial analyses, we also employed exploratory factor analysis (Tucker and MacCallum 1997; Fabrigar et al. 1999) to determine item dimensionality of larger item pools.

The use of multidimensional models also allows an assessment of the estimated correlation(s) between latent factors, which provide(s) information on the similarity of the different dimensions measured by related item sets.

Generally, maximum likelihood estimation and covariance matrices are not appropriate for analyses of (categorical) questionnaire items because the approach treats items as if they were continuous. Therefore, the analyses of ICILS 2018 relied on robust weighted least squares estimation (see Muthén et al. 1997; Flora and Curran 2004) to estimate the confirmatory factor models. The software package used for the estimations was *MPLUS 7* (Muthén and Muthén 2012).

Confirmatory factor analyses were carried out for sets of conceptually related questionnaire items that measured between one or more different dimensions. This approach allowed an assessment of the measurement model as well as of the associations between related latent factors. The scaling analyses were restricted to data from those countries that met sample participation requirements (see Chapter 8 for further information). National samples of students, teachers, and schools received weights that ensured equal representations of countries in the analyses.

In international studies, model parameters may vary across country and it may not be appropriate to assume the same factor structure for each population. To test parameter invariance, multiple-group modeling, as an extension of CFA, offers an approach to test the equivalence of measurement models across sub-samples (Little 1997; Byrne 2008).

When considering a model where respondents belong to different groups indexed as g = 1, 2, ..., G, the multiple-group factor model becomes

$$x_g = \Lambda_{xg}\xi_g + \delta_g$$

A test of factorial invariance (H_Λ) where factor loadings are defined as being equal (often referred to as "metric equivalence") (Horn and McArdle 1992) can be defined as

$$H_\Lambda : \Lambda_1 = \Lambda_1 = \Lambda_2 = \ldots = \Lambda_g$$

Typically, model-fit indices are compared across different multiple-group models, each with an increasing degree of constraints; from relaxed models with no constraints through to constrained models with largely invariant model parameters.

In this report, for all student and teacher questionnaire scales,[7] three different multiple-group models for CFA were estimated with different levels of constraints on parameters:

A. Unconstrained models where all parameters are estimated as country-specific (*configural invariance*);
B. Models with constrained factor loadings across countries (*metric invariance*); and
C. Models with constraints on factor loadings **and** intercepts (*scalar invariance*).

Models with confirmed scalar invariance are the only ones that ensure full comparability of measurement models across participating countries. When comparing model fit across the three conditions, it needs to be acknowledged that with data from large samples, as is typically the case in international large-scale assessments, even very small differences appear to be significant. This makes hypothesis testing using tests of statistical significance difficult and therefore, when reviewing results, it is more appropriate to focus on relative changes in model fit across the three models with different levels of constraints.

7 For school questionnaire data, we did not conduct this type of analyses given the relatively small size of national samples.

Item response modeling

In line with the scaling of test item data (see Chapter 11), item response modeling was used to scale questionnaire items. The one-parameter (*Rasch*) model (Rasch 1960) for dichotomous items models the probability of selecting an item category 1 instead of 0 as:

$$P_i(\theta_n) = \frac{exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

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.

In the case of items with more than two categories (as, for example, with Likert-type items), this model can be generalized to the (Rasch) partial credit model (Masters and Wright 1997), which takes the form of:

$$P_{x_i}(\theta_n) = \frac{exp \sum_{j=0}^{x}(\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} exp \sum_{j=0}^{h}(\theta_n - \delta_i + \tau_{ij})} \quad x_i = 0,1,\ldots,m_i$$

where $P_{Xi}(\theta_n)$ 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.

Weighted mean-square statistics (*infit*), statistics based on model residuals, were used in conjunction with a wide range of further item statistics to assess the fit of the IRT model. ICILS 2018 used the ACER ConQuest software package (Adams et al. 2015) for the analysis of item scaling properties and the estimation of item parameters.

The international item parameters were derived using equally weighted national datasets[8]:

A *Calibration of item parameters for the student questionnaire*: This was done based on a pooled database with equally weighted national samples from 11 countries that met sample participation requirements for the student survey.

B *Calibration of item parameters for the teacher questionnaire*: This was done based on a pooled database with equally weighted national samples from seven countries that met sample participation requirements for the teacher survey.

C *Calibration of item parameters for school questionnaire*: This was done based on a pooled database with equally weighted national samples from 11 countries that met sample participation requirements for the student survey.

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:

$$\sum_{i \in \Omega} \left[\left(r_x + \frac{J_n}{2I_n} \right) - \sum_{i=1}^{k} \frac{exp(\sum_{j=0}^{x} \theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} exp \sum_{j=0}^{h}(\theta_n - \delta_i + \tau_{ij})} \right] = 0$$

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 θ) is used as a weight function to account for the bias inherent in maximum likelihood estimation (see Warm 1989). We used the ACER ConQuest software for the pre-calibration of item parameters in order to subsequently derive scale scores.

8 Data from benchmarking participants were not included in the item parameter calibrations.

For all questionnaire scales in ICILS 2018 the transformation of weighted likelihood estimates to an international metric resulted in reporting scales with an ICILS 2018 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:

$$\theta_n' = 50 + 10\left(\frac{\theta_n - \mu_{\theta(ICILS18)}}{\sigma_{\theta(ICILS18)}}\right)$$

where θ_n' are the scores in the international metric, θ_n are the original weighted likelihood estimates in logits, and $\mu_{\theta(ICILS18)}$ is the international mean of logit scores with equally weighted country subsamples. $\sigma_{\theta(ICILS18)}$ is the corresponding international standard deviation of the original weighted likelihood estimates.

Table 12.1 presents the means and standard deviations used to transform the original scale scores for the student, teacher, ICT coordinator, and school principal questionnaires into the international metric.

Table 12.1: Transformation parameters for new ICILS 2018 questionnaire scales (means and standard deviations of original IRT logit scores)

International student questionnaire		
Scale	Mean	SD
S_GENACT	-0.63	1.49
S_SPECACT	-0.89	0.97
S_USECOM	0.43	0.88
S_USEINF	-1.23	0.97
S_ACCONT	0.26	1.05
S_USESTD	-0.51	1.03
S_SPECLASS	-2.3	1.69
S_GENCLASS	-0.97	1.95
S_ICTLRN	0.66	1.69
S_GENEFF	1.86	1.36
S_SPECEFF	-0.1	1.61
S_ICTPOS	1.61	1.79
S_ICTNEG	0.66	1.26
S_ICTFUT	0.25	2.00
S_CODLRN	0.25	1.85
ICT coordinator questionnaire		
Scale	Mean	SD
C_HINPED	0.23	1.41
C_HINRES	-0.35	1.48

Teacher questionnaire		
Scale	Mean	SD
T_CLASACT	-0.47	1.88
T_CODEMP	0.36	1.83
T_COLICT	1.00	2.59
T_ICTEFF	2.32	1.47
T_ICTEMP	1.00	1.94
T_ICTPRAC	-0.56	1.80
T_PROFREC	-0.41	1.19
T_PROFSTR	-1.16	1.55
T_RESRC	0.19	1.90
T_USETOOL	-1.85	1.43
T_USEUTIL	-0.23	1.47
T_VWNEG	0.08	1.57
T_VWPOS	1.78	2.13
School principal questionnaire		
Scale	Mean	SD
P_EXPLRN	0.90	1.90
P_EXPTCH	1.32	2.33
P_ICTCOM	0.97	1.47
P_ICTUSE	0.05	0.94
P_PRIORH	1.54	1.78
P_PRIORS	1.52	1.64
P_VWICT	3.63	2.24

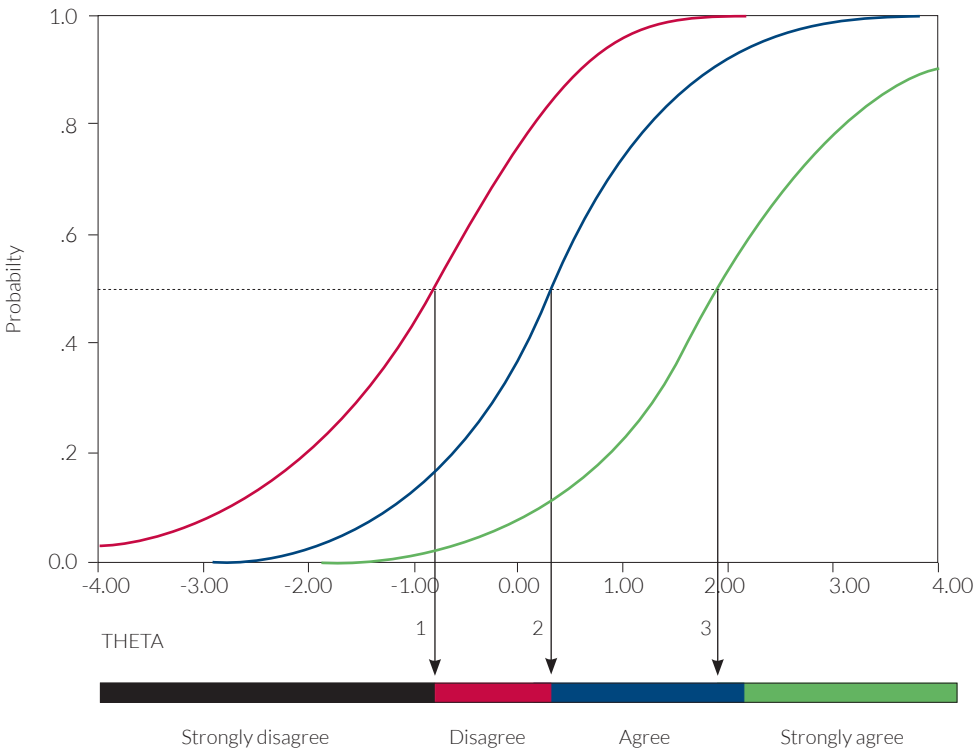
Describing questionnaire scale indices

Questionnaire scales derived from weighted likelihood estimates (logits) present values on a continuum with an ICILS 2018 average of 50 and a standard deviation of 10 (for equally weighted national samples). This allows an interpretation of these scores by comparing individual scores or group average scores with the ICILS 2018 average but the individual scores do not reveal anything about the actual item responses and the extent to which respondents endorsed the items used to measure the latent variable. The scaling model used to derive individual scores allows the development of descriptions of these scales through a mapping of scale scores to (expected) item responses.[9]

It is possible to describe item characteristics by using the parameters of the partial credit model to provide an estimate for each category of its probability of being chosen relative to the probabilities of all higher categories. This process is equivalent to computing the odds of scoring higher than a particular category.

Figure 12.1 presents the results of plotting these cumulative probabilities against scale scores for a fictitious item, where respondents rate their agreement or a disagreement to a statement on a four-point scale. The three vertical lines denote those points on the latent continuum where it becomes more likely to score > 0, > 1, or > 2. These locations Γ_k are *Thurstonian* thresholds which can be obtained through an iterative procedure that calculates summed probabilities for each category at each (decimal) point on the latent variable.

Figure 12.1: Summed category probabilities for fictitious item

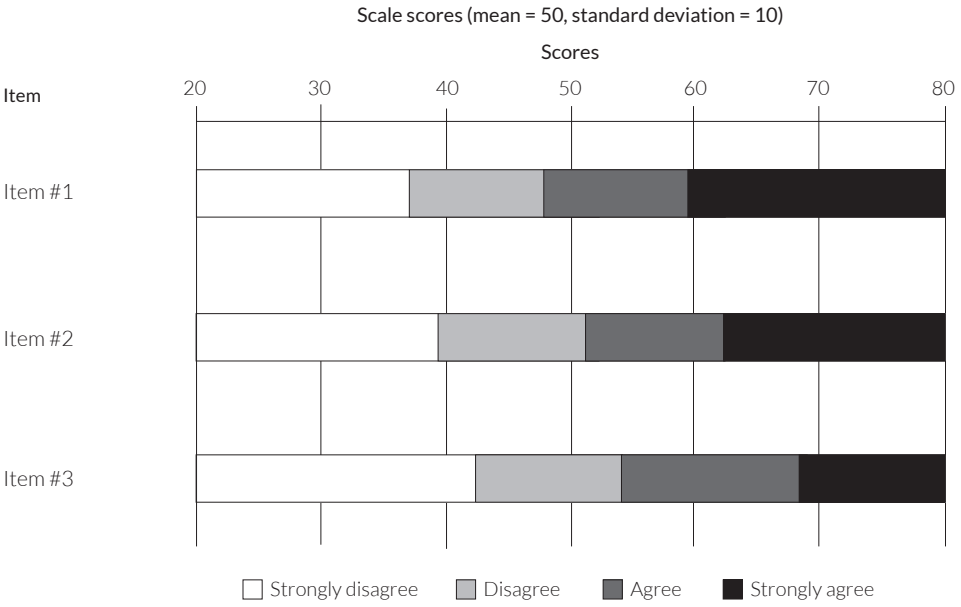


9 This approach was also used in the IEA ICCS 2009 and 2016 surveys (see Schulz and Friedman 2011, 2018) and the ICILS 2013 survey (see Schulz and Friedman 2015).

Summed probabilities are not identical to expected item scores and have to be understood in terms of the probability of scoring *at least* a particular category.[10] Thurstonian thresholds can be used to indicate for each item category those points on a scale at which respondents have a 0.5 probability of scoring in this category or higher. For example, in the case of Likert-type items with the categories *strongly disagree*, *disagree*, *agree* and *strongly agree*, we can determine at what point of a scale a respondent has a 50 percent likelihood of agreeing (or strongly agreeing) with the item.

The item-by-score maps included in ICILS 2018 reports predict the minimum coded score (e.g., 0 = "strongly disagree," 1 = "disagree," 2 = "agree," and 3 = "strongly agree") a respondent would obtain on a Likert-type item. For example, we could predict that students with a certain scale score would have a 50 percent probability of agreeing (or strongly agreeing) with a particular item (see the example item-by-score map in Figure 12.2). For each item, it is thus possible to determine Thurstonian thresholds: the points at which a minimum item score becomes more likely than any lower score to occur and which determine the boundaries between item categories on the item-by-score map.

Figure 12.2: Example of questionnaire item-by-score map



Example of how to interpret the item-by-score map

- #1: A respondent with score 30 has more than 50% probability to strongly disagree with all three items
- #2: A respondent with score 40 has more than 50% probability **not** to strongly disagree with items 1 and 2 but to strongly disagree with item 3
- #3: A respondent with score 50 has more than 50% probability to agree with items 1 and to disagree with items 2 and 3
- #4: A respondent with score 60 has more than 50% probability to strongly agree with items 1 and to at least agree with items 2 and 3
- #5: A respondent with score 70 has more than 50% probability to strongly agree with items 1, 2, and 3

10 Other ways of describing item characteristics based on the partial credit model are item characteristic curves, which involve plotting the individual category probabilities and the expected item score curves (for a detailed description, see Masters and Wright 1997).

This information can also be summarized by calculating the average thresholds across all items in a scale. For example, it is possible to do this for the second threshold of a four-point Likert-type scale, which allows the prediction of how likely it would be for a respondent with a certain scale score to have responses in the two lower or upper categories (on average across items). For ICILS 2018 we used this approach in the case of items measuring agreement to distinguish between scale scores for respondents who were most likely to agree or disagree with the "average item" used for measuring the respective latent trait.

In the reporting tables for questionnaire scales, we depicted national average student or teacher scale scores as boxes that indicated their mean values plus/minus sampling error and that were set in graphical displays featuring two underlying colors. National average scores located in the darker shaded area indicated that, on average across items, student or teacher responses had resided in the lower item categories ("disagree or strongly disagree" or "less than once a week"). If these scores were found in the lighter shaded area, however, then students' or teachers' average item responses would have been in the upper item response categories ("agree or strongly agree" or "at least once a week").

Scaled indices

Student questionnaire

National index of students' socioeconomic background

The multivariate analyses presented in the international report (Fraillon et al. 2020) include a composite index reflecting students' socioeconomic background. The *national index of students' socioeconomic background* (S_NISB) was 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). For the S_HISCED index, we collapsed the lowest two categories to have an indicator variable with four categories: lower-secondary or below, upper-secondary, tertiary non-university, and university education. The S_HOMLIT index was reduced from five to four categories (0 to 10 books; 11 to 25 books; 26 to 100 books; more than 100 books) collapsing the two highest categories. This was done for both indices on parental education and home literacy as prior analyses had shown approximately linear associations across these categories with CIL and CT scores as well as other indicators of socioeconomic background.

In order to impute 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.

After converting the resulting variables including the imputed values into z-standardized variables (with a mean of 0 and a standard deviation of 1 for each national dataset), principal component analysis of these indicator variables were conducted separately for each weighted national sample.

The final S_NISB scores consists of factor scores for the first principal component with national averages of 0 and national standard deviations of 1. Table 12.2 shows the factor loadings and reliabilities for each national sample.

Students' use of ICT for different activities

Question 19 of the student questionnaire asked students to indicate their use of ICT for a range of different activities. For each of the eight activities, they were asked to select "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 items in this question were used to derive two scales. The first one reflected *students' use of general applications for activities* (S_GENACT) as based on the first three items of the question and had an average reliability of 0.70 across the national samples, with Cronbach's alpha coefficients ranging from 0.63 to 0.84 (see Table 12.3). The second scale reflected *students' use of*

Table 12.2: Factor loadings and reliabilities for the national index of students' socioeconomic background

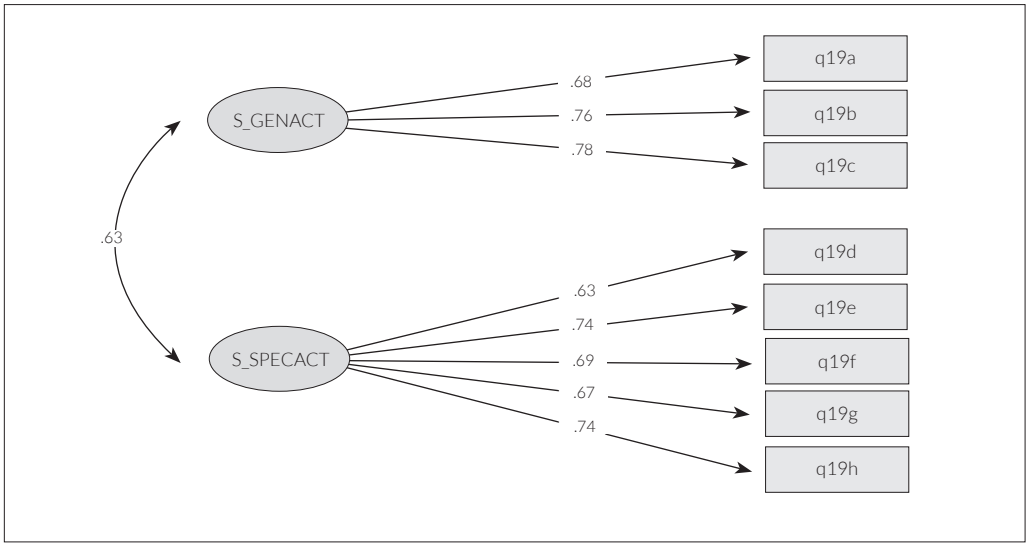
Country	Factor loadings for:			Cronbach's alpha
	Highest parental occupation	Highest parental education	Books at home	
Chile	0.90	0.88	0.70	0.77
Denmark	0.79	0.79	0.72	0.65
Finland	0.80	0.76	0.56	0.52
France	0.80	0.78	0.69	0.63
Germany	0.82	0.72	0.66	0.61
Italy	0.82	0.82	0.72	0.70
Kazakhstan	0.78	0.78	0.57	0.51
Korea, Republic of	0.72	0.75	0.64	0.51
Luxembourg	0.81	0.79	0.75	0.67
Moscow (Russian Federation)	0.76	0.78	0.64	0.55
North Rhine-Westphalia (Germany)	0.81	0.68	0.69	0.59
Portugal	0.84	0.86	0.74	0.74
United States	0.80	0.83	0.68	0.66
Uruguay	0.82	0.82	0.75	0.71
ICILS 2018 average	0.80	0.78	0.67	0.62

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

specialist applications for activities (S_SPECACT). It was derived from the remaining five items in the question with reliabilities ranging from 0.65 to 0.82, with an average of 0.73. The higher values on each scale reflect more frequent use of ICT for the corresponding activities.

Figure 12.3 depicts the results from the confirmatory factor analysis of the scaled items from the question. The model had satisfactory fit for the two-factor model, and we found a high positive correlation between the two latent factors. When reviewing measurement invariance using multiple-group models with different constraints, the model fit changed only marginally which indicates a relatively high degree of invariance for this model. The reliabilities were satisfactory for all countries. The item parameters for both scales that were used to derive the IRT scale scores are presented in Table 12.4.

Figure 12.3: Confirmatory factor analysis of items measuring students' use of ICT for activities



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.060	0.077	0.096	0.091
CFI	0.98	0.97	0.93	0.89
TLI	0.96	0.95	0.92	0.93

Table 12.3: Reliabilities for scales measuring students' participation in out-of-school activities

Country	Scale reliability (Cronbach's alpha)	
	S_GENACT	S_SPECACT
Chile	0.68	0.70
Denmark	0.70	0.74
Finland	0.71	0.73
France	0.65	0.71
Germany	0.69	0.70
Italy	0.63	0.71
Kazakhstan	0.75	0.78
Korea, Republic of	0.84	0.82
Luxembourg	0.68	0.75
Moscow (Russian Federation)	0.72	0.75
North Rhine-Westphalia (Germany)	0.69	0.65
Portugal	0.69	0.73
United States	0.67	0.70
Uruguay	0.71	0.70
ICILS 2018 average	0.70	0.73

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.4: Item parameters for scales measuring students' use of ICT for activities

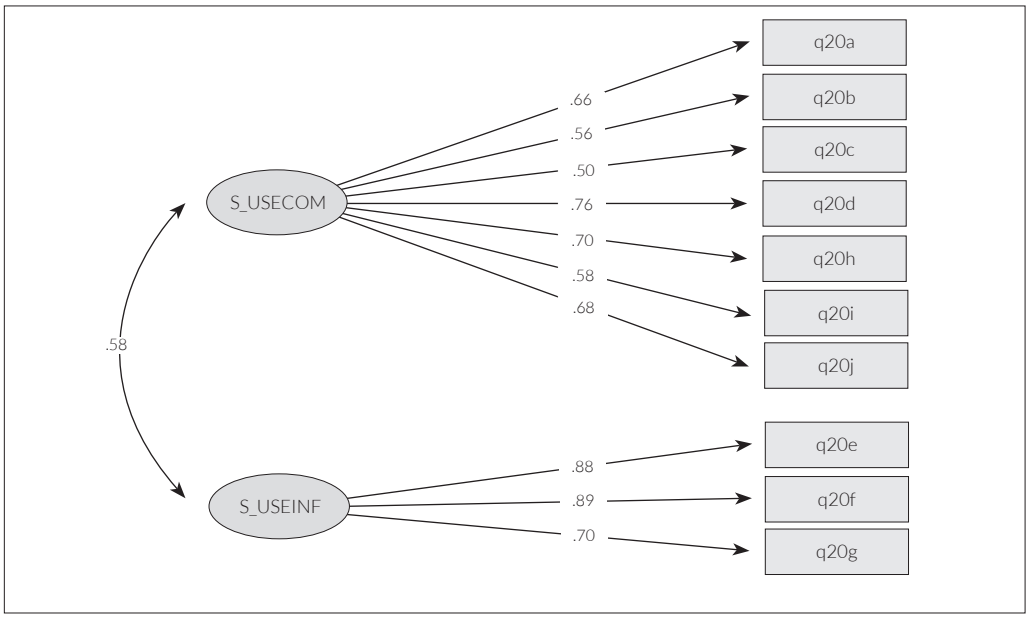
Scale or item	Question/item wording	Item parameters				
		Delta	Tau(1)	Tau(2)	Tau(3)	Tau(4)
S_GENACT	*How often do you use ICT for each of the following activities?*					
IS2G19A	Write or edit documents	-0.42	-1.89	-0.57	0.34	2.12
IS2G19B	Use a spreadsheet to do calculations, store data or plot graphs (e.g. using [Microsoft Excel ®])	0.33	-1.61	-0.46	0.00	2.08
IS2G19C	Create a simple "slideshow" presentation (e.g. using [Microsoft PowerPoint ®])	0.10	-2.61	-0.71	0.67	2.65
S_SPECACT	*How often do you use ICT for each of the following activities?*					
IS2G19D	Record or edit videos	-0.44	-0.78	-0.17	-0.09	1.03
IS2G19E	Write computer programs, scripts or apps (e.g. using [Logo, LUA, or Scratch])	0.29	-0.49	-0.22	-0.19	0.90
IS2G19F	Use drawing, painting or graphics software or [apps]	-0.18	-0.64	-0.05	-0.03	0.71
IS2G19G	Produce or edit music	-0.14	0.34	-0.13	-0.22	0.01
IS2G19H	Build or edit a webpage	0.46	0.16	-0.20	-0.26	0.30

Students' use of ICT for communication activities

In Question 20, respondents were asked to indicate the frequency with which they use ICT for 10 different communication activities. For each activity, they were asked to select "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." Two scales were derived from this item set. The first set reflected *students' use of ICT for social communication* (S_USECOM) while the second reflected *students' use of ICT for exchanging information* (S_USEINF). Higher scores on both of these scales indicated more frequent use of ICT for these purposes.

Figure 12.4 depicts the results from the confirmatory factor analysis of the scaled items from the question. There was only marginally satisfactory fit for the two-factor model and there was a moderate correlation between the two latent factors (r=0.58). When reviewing measurement invariance using across multiple-group models with different constraints, the model fit was not satisfactory for the most constrained models, which suggests a lack of measurement invariance across countries. The national scale reliabilities for the two scales are presented in Table 12.5. Both scales had an acceptable average reliability across participating countries (0.78 and 0.75 respectively), with national reliabilities ranging 0.72 to 0.82 for S_USECOM and 0.72 to 0.80 for S_USEINF. The item parameters for the two scales are presented in Table 12.6.

Figure 12.4: Confirmatory factor analysis of items measuring students' use of ICT for communication activities



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.090	0.107	0.109	0.108
CFI	0.93	0.92	0.90	0.84
TLI	0.91	0.90	0.89	0.89

Table 12.5: Reliabilities for scales measuring students' use of ICT for communication activities

Country	Scale reliability (Cronbach's alpha)	
	S_USECOM	S_USEINF
Chile	0.80	0.73
Denmark	0.73	0.75
Finland	0.74	0.73
France	0.80	0.73
Germany	0.73	0.76
Italy	0.77	0.74
Kazakhstan	0.82	0.77
Korea, Republic of	0.81	0.80
Luxembourg	0.76	0.80
Moscow (Russian Federation)	0.76	0.75
North Rhine-Westphalia (Germany)	0.72	0.73
Portugal	0.76	0.76
United States	0.82	0.76
Uruguay	0.78	0.72
ICILS 2018 average	0.77	0.75

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

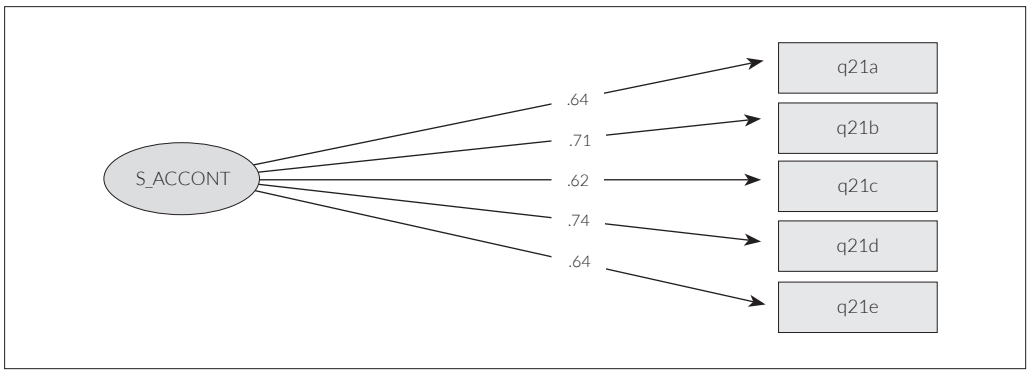
Table 12.6: Item parameters for scales measuring students' use of ICT for communication activities

Scale or item	Question/item wording	Item parameters				
		Delta	Tau(1)	Tau(2)	Tau(3)	Tau(4)
S_USECOM	*How often do you use ICT to do each of the following communication activities?*					
IS2G20A	Share news about current events on social media	0.51	-0.14	0.10	-0.30	0.34
IS2G20B	Communicate with friends, family, or other people using instant messaging, voice or video chat (e.g. [Skype, WhatsApp, Viber])	-0.78	0.06	0.34	-0.10	-0.30
IS2G20C	Send texts or instant messages to friends, family, or other people	-0.71	0.15	0.38	-0.15	-0.38
IS2G20D	Write posts and updates about what happens in your life on social media	0.79	-0.09	-0.13	-0.26	0.48
IS2G20H	Post images or video in social networks or online communities (e.g. [Facebook, Instagram or YouTube])	0.44	-0.49	-0.21	0.04	0.67
IS2G20I	Watch videos or images that other people have posted online	-0.78	0.12	0.34	-0.12	-0.34
IS2G20J	Send or forward information about events or activities to other people	0.52	-0.37	-0.20	-0.09	0.66
S_USEINF	*How often do you use ICT to do each of the following communication activities?*					
IS2G20E	Ask questions on forums or [Q&A, question and answer] websites	-0.11	-0.55	-0.49	0.08	0.97
IS2G20F	Answer other peoples' questions on forums or [Q&A, question and answer] websites	-0.10	-0.37	-0.43	0.03	0.77
IS2G20G	Write posts for your own blog (e.g. [WordPress, Tumblr, Blogger])	0.21	0.34	-0.49	-0.28	0.43

Students' use of ICT for accessing content from the internet

Students are asked in Question 21 to indicate the frequency with which they use ICT for a list of leisure activities ("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 first five items of the question were used to derive a scale of *students' use of ICT for accessing content from the internet* (S_ACCONT). Higher scores on the scale correspond to more frequent use of ICT for leisure activities. A confirmatory factor analysis assuming a one-dimensional model for the five items revealed only marginally satisfactory fit (see Figure 12.5). A review of measurement invariance indicated some variation in model fit across multiple-group models with different constraints that suggests considerable variation in measurement characteristics. The reliabilities (Cronbach's alpha) of the scale were satisfactory for all countries with an average of 0.76 across countries (Table 12.7). The item parameters for the scale are presented in Table 12.8.

Figure 12.5: Confirmatory factor analysis of items measuring students' use of ICT for leisure activities



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.098	0.102	0.079	0.094
CFI	0.97	0.97	0.97	0.90
TLI	0.95	0.95	0.97	0.95

Table 12.7: Reliabilities for scale measuring students' use of ICT for leisure activities

Country	Scale reliability (Cronbach's alpha)
	S_ACCONT
Chile	0.76
Denmark	0.73
Finland	0.78
France	0.76
Germany	0.72
Italy	0.73
Kazakhstan	0.79
Korea, Republic of	0.83
Luxembourg	0.74
Moscow (Russian Federation)	0.68
North Rhine-Westphalia (Germany)	0.73
Portugal	0.74
United States	0.79
Uruguay	0.76
ICILS 2018 average	0.75

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.8: Item parameters for scale measuring students' use of ICT for leisure activities

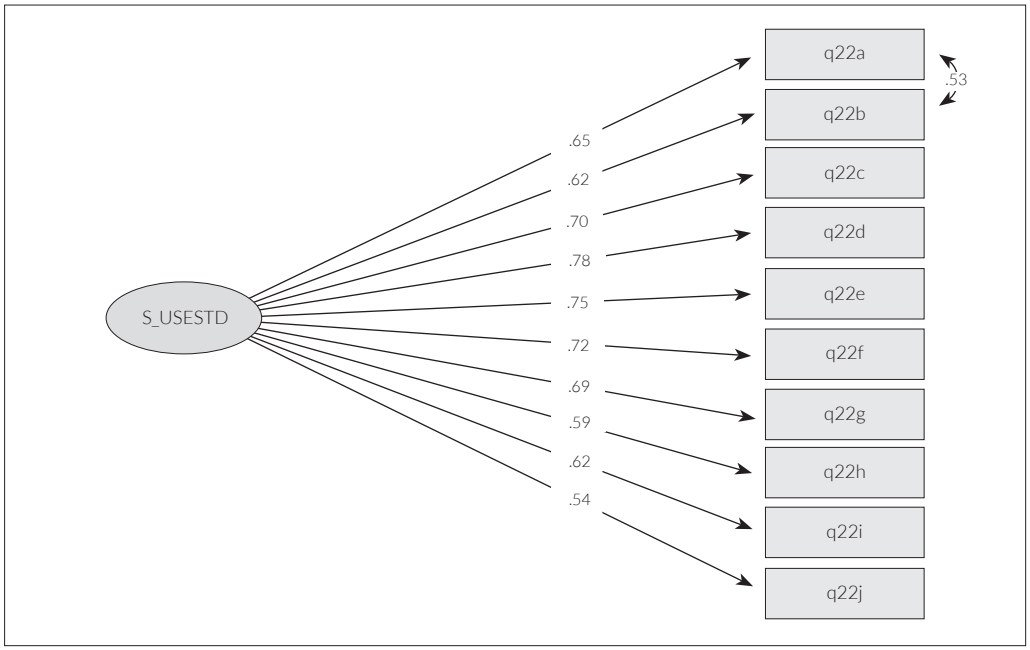
Scale or item	Question/item wording	Item parameters				
		Delta	Tau(1)	Tau(2)	Tau(3)	Tau(4)
S_ACCONT	*How often do you use ICT to do each of the following leisure activities?*					
IS2G21A	Search the Internet to find information about places to go or activities to do	0.29	-1.26	-0.25	0.17	1.34
IS2G21B	Read reviews on the Internet of things you might want to buy	0.30	-0.82	-0.41	0.05	1.19
IS2G21C	Read news stories on the Internet	0.00	-0.57	-0.09	-0.03	0.69
IS2G21D	Search for online information about things you are interested in	-0.60	-0.66	-0.31	-0.09	1.06
IS2G21E	Use websites, forums, or online videos to find out how to do something	0.01	-0.65	-0.46	-0.04	1.15

Students' use of ICT for study purposes

In Question 22, students were asked to indicate the frequency that they use ICT for 10 items on different school-related purposes. For each item, they were asked to select "never," "less than once a month," "at least once a month but not every week," "at least once a week but not every school day," or "every school day." All items in the question were used to derive a scale of *students' use of ICT for study purposes* (S_USESTD), where higher scale scores reflect more frequent use of ICT. The results from a confirmatory factor analysis for the model are presented in Figure 12.6.

Overall, the model showed only marginally satisfactory fit, and reviews of multiple-group model comparisons show that there were differences in model fit between the metric and scalar invariance model that suggest there was some variation in measurement characteristics. The reliabilities (Cronbach's alpha) of the scale are presented in Table 12.9 and show a high average reliability of 0.83, ranging between 0.77 and 0.89 across countries. The item parameters for the scale are presented in Table 12.10.

Figure 12.6: Confirmatory factor analysis of items measuring students' use of ICT for study purposes



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.072	0.086	0.097	0.108
CFI	0.97	0.96	0.93	0.86
TLI	0.96	0.94	0.92	0.91

Table 12.9: Reliabilities for scale measuring students' use of ICT for study purposes

Country	Scale reliability (Cronbach's alpha)
	S_USESTD
Chile	0.82
Denmark	0.77
Finland	0.86
France	0.81
Germany	0.82
Italy	0.80
Kazakhstan	0.87
Korea, Republic of	0.89
Luxembourg	0.84
Moscow (Russian Federation)	0.82
North Rhine-Westphalia (Germany)	0.81
Portugal	0.84
United States	0.84
Uruguay	0.84
ICILS 2018 average	0.83

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.10: Item parameters for scale measuring students' use of ICT for study purposes

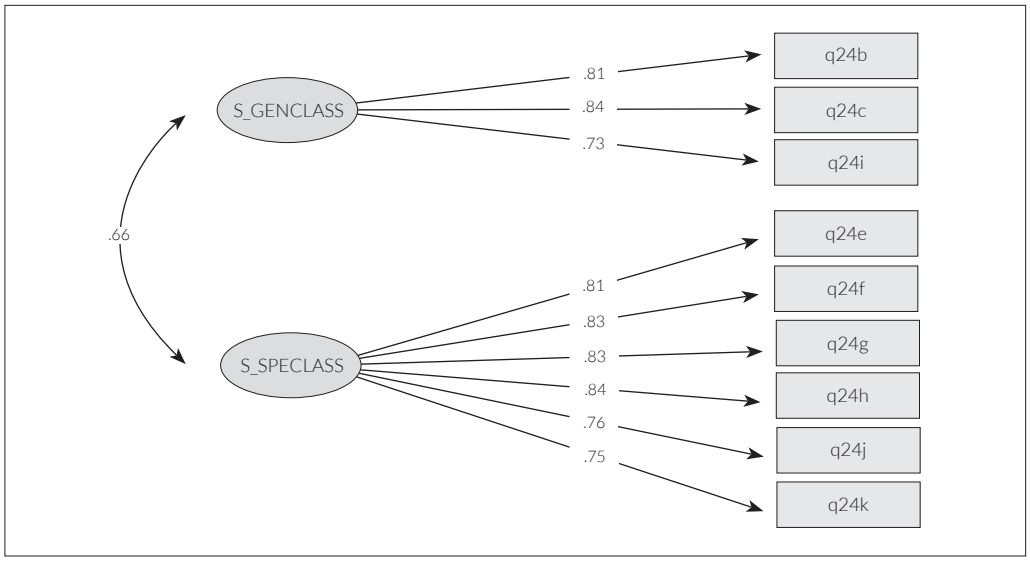
Scale or item	Question/item wording	Item parameters				
		Delta	Tau(1)	Tau(2)	Tau(3)	Tau(4)
S_USESTD	*How often do you use ICT for the following school-related purposes?*					
IS2G22A	Prepare reports or essays	-0.08	-1.38	-0.51	0.23	1.66
IS2G22B	Prepare presentations	-0.06	-1.87	-0.55	0.49	1.93
IS2G22C	Work online with other students	0.08	-0.46	-0.26	-0.11	0.83
IS2G22D	Complete [worksheets] or exercises	-0.09	-0.60	-0.22	-0.11	0.94
IS2G22E	Organize your time and work	-0.03	-0.10	-0.24	-0.13	0.47
IS2G22F	Take tests	0.28	-0.71	-0.48	-0.07	1.27
IS2G22G	Use software or applications to learn skills or a subject (e.g. mathematics tutoring software, language learning software)	0.11	-0.54	-0.40	-0.08	1.03
IS2G22H	Use the Internet to do research	-1.12	-0.92	-0.36	0.20	1.08
IS2G22I	Use coding software to complete assignments (e.g. [Scratch])	0.57	-0.25	-0.46	-0.15	0.86
IS2G22J	Make video or audio productions	0.33	-0.17	-0.05	-0.12	0.34

Students' use of ICT for class activities

Question 24 required students to indicate how often they use different ICT-related tools during class (selecting from "never," "in some lessons," "in most lessons," or "in every or almost every lesson"). Three of the items were used to derive a scale on *students' use of general applications in class* (S_GENCLASS) and six of the items were used to derive a scale of *students' use of specialist applications in class* (S_SPECLASS). Higher scale scores reflect more frequent use of the respective type of ICT applications for class activities.

A confirmatory factor analysis using all scaled items from this question (see Figure 12.7) revealed satisfactory fit for a two-dimensional model with a strong positive correlation between the two latent factors (0.66). The fit of the model was also acceptable when comparing multiple-group models with different constraints, which suggests a relatively high degree of measurement invariance. The national reliabilities for the two scales are presented in Table 12.11. Both scales had acceptable reliabilities (Cronbach's alpha) with an average of 0.72 and 0.84 across countries respectively (ranging from 0.53 to 0.81 for S_GENCLASS and ranging from 0.78 to 0.92 for S_SPECLASS). The item parameters for both scales are displayed in Table 12.12.

Figure 12.7: Confirmatory factor analysis of items measuring students' report on the use of ICT for class activities



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.075	0.087	0.096	0.085
CFI	0.97	0.97	0.96	0.96
TLI	0.97	0.96	0.96	0.97

Table 12.11: Reliabilities for scales measuring students' reports on the use of ICT for class activities

Country	Scale reliability (Cronbach's alpha)	
	S_GENCLASS	S_SPECLASS
Chile	0.76	0.84
Denmark	0.53	0.82
Finland	0.70	0.78
France	0.73	0.78
Germany	0.67	0.82
Italy	0.74	0.82
Kazakhstan	0.76	0.87
Korea, Republic of	0.81	0.92
Luxembourg	0.76	0.89
Moscow (Russian Federation)	0.73	0.82
North Rhine-Westphalia (Germany)	0.69	0.84
Portugal	0.78	0.86
United States	0.70	0.85
Uruguay	0.72	0.86
ICILS 2018 average	0.72	0.84

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.12: Item parameters for scales measuring students' reports on the use of ICT for class activities

Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
S_GENCLASS	*When studying throughout this school year, how often did you use the following tools during class?*				
IS2G24B	Word-processing software (e.g. [Microsoft Word ®])	-0.14	-2.51	0.61	1.90
IS2G24C	Presentation software (e.g. [Microsoft PowerPoint ®])	0.03	-2.98	0.53	2.45
IS2G24I	Computer-based information resources (e.g. websites, wikis, encyclopaedia)	0.12	-2.26	0.17	2.08
S_SPECLASS	*When studying throughout this school year, how often did you use the following tools during class?*				
IS2G24E	Multimedia production tools (e.g. media capture and editing, web production)	-0.07	-1.86	0.39	1.47
IS2G24F	Concept mapping software (e.g. [Inspiration ®], [Webspiration ®])	0.37	-1.67	0.19	1.48
IS2G24G	Tools that capture real-world data (e.g. speed, temperature) digitally for analysis	0.16	-1.82	0.33	1.49
IS2G24H	Simulations and modelling software	0.35	-1.68	0.26	1.42
IS2G24J	Interactive digital learning resources (e.g. learning games or applications)	-0.48	-2.06	0.45	1.61
IS2G24K	Graphing or drawing software	-0.33	-1.82	0.31	1.51

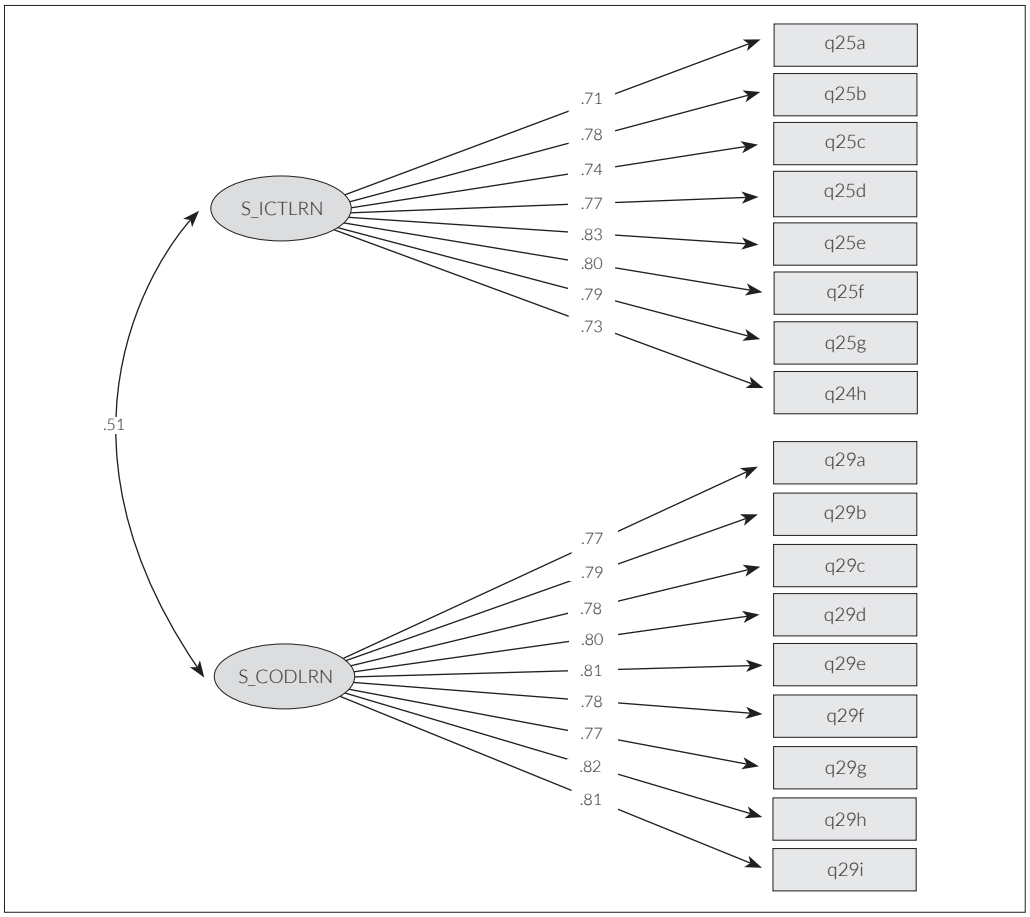
Students' perceptions of school learning of ICT and coding tasks

In Question 25, students were asked to indicate the extent to which they had learnt about eight different tasks at school that are related to using ICT. They had to select between four different options for each task ("to a large extent," "to a moderate extent," "to a small extent," or "not at all"). The items were used to derive a scale of *students' learning of ICT tasks at school* (S_ICTLRN). Higher scale scores corresponded to higher perceived learning of different tasks involving ICT.

Question 29 required students to indicate the extent to which they had been taught how to do tasks that are related to coding (selecting from the same response options as in Question 25). The nine items in the scale were used to derive a scale of *students' learning of ICT coding tasks at school* (S_CODLRN).

Figure 12.8 shows the results from a confirmatory factor analysis of all scaled items measuring students' perceptions of school learning of ICT and coding tasks. The model fit was satisfactory and a moderate correlation was observed between the two latent factors (0.51). When reviewing measurement invariance using multiple-group models with different constraints, the model fit changed only marginally which indicates a relatively high degree of invariance for this model.

Figure 12.8: Confirmatory factor analysis of items measuring students' perceptions of school learning of ICT and coding tasks



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.063	0.071	N/A	0.078
CFI	0.97	0.97	N/A	0.94
TLI	0.96	0.96	N/A	0.95

Table 12.13 shows the scale reliabilities (Cronbach's alpha) for the two scales reflecting the extent of learning of ICT and coding tasks. The reliabilities were satisfactory for all countries. The item parameters for both scales that were used to derive the IRT scale scores are presented in Table 12.14. S_ICTLRN had a high level of reliability (Cronbach's alpha) with an average reliability of 0.88 across countries (ranging from 0.83 to 0.94), and S_CODLRN also had a high average reliability (Cronbach's alpha = 0.90) across countries (ranging from 0.85 to 0.96).

Table 12.13: Reliabilities for scales measuring students' perceptions of school learning of ICT and coding tasks

Country	Scale reliability (Cronbach's alpha)	
	S_ICTLRN	S_CODLRN
Chile	0.86	0.90
Denmark	0.83	0.85
Finland	0.92	0.91
France	0.84	0.88
Germany	0.86	0.89
Italy	0.87	0.88
Kazakhstan	0.89	0.90
Korea, Republic of	0.94	0.96
Luxembourg	0.88	0.91
Moscow (Russian Federation)	0.94	0.91
North Rhine-Westphalia (Germany)	0.87	0.89
Portugal	0.91	0.92
United States	0.89	0.90
Uruguay	0.87	0.90
ICILS 2018 average	0.88	0.90

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.14: Item parameters for scales reports on the use of ICT for class activities

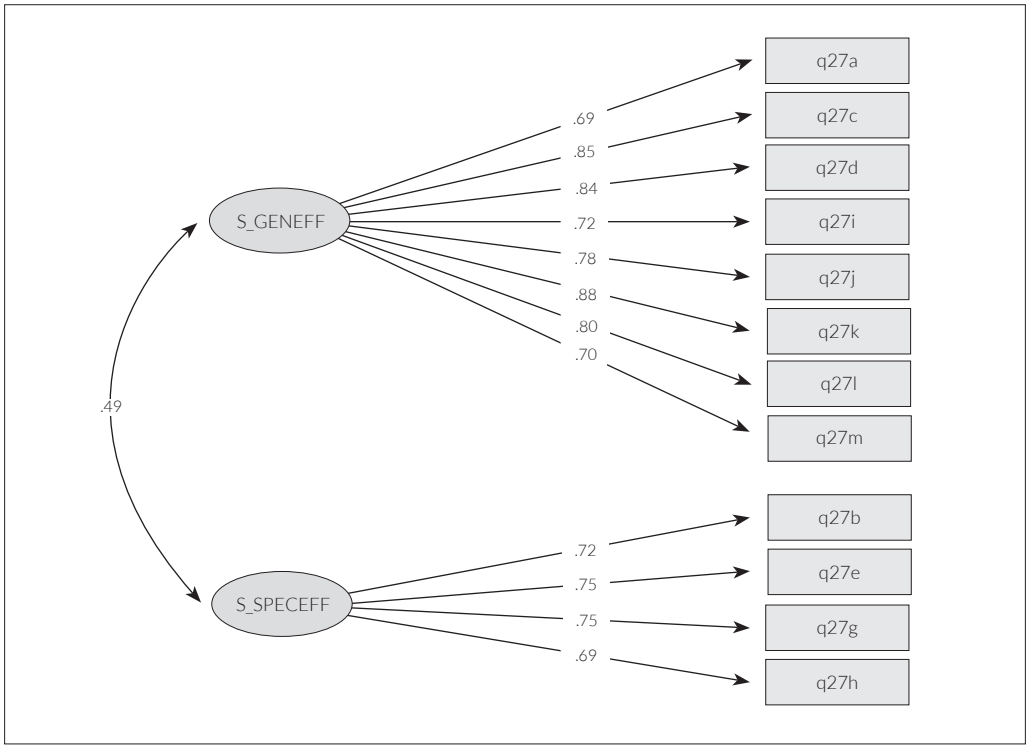
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
S_ICTLRN	*At school, to what extent have you learned how to do the following tasks?*				
IS2G25A	Provide references to Internet sources	-0.04	-1.57	-0.32	1.88
IS2G25B	Search for information using ICT	-0.38	-1.26	-0.34	1.60
IS2G25C	Present information for a given audience or purpose using ICT	0.10	-1.35	-0.38	1.73
IS2G25D	Work out whether to trust information from the Internet	0.08	-1.36	-0.30	1.66
IS2G25E	Decide what information obtained from the Internet is relevant to include in school work	-0.03	-1.38	-0.37	1.75
IS2G25F	Organize information obtained from Internet sources	-0.04	-1.50	-0.33	1.82
IS2G25G	Decide where to look for information on the Internet about an unfamiliar topic	-0.01	-1.32	-0.32	1.64
IS2G25H	Use ICT to collaborate with others	0.31	-1.12	-0.27	1.39
S_CODLRN	*When studying during the current school year, to what extent have you been taught how to do the following tasks?*				
IS2G29A	To display information in different ways	-0.89	-1.59	-0.49	2.08
IS2G29B	To break a complex process into smaller parts	0.20	-1.56	-0.51	2.06
IS2G29C	To understand diagrams that describe or show real-world problems	-0.30	-1.64	-0.29	1.93
IS2G29D	To plan tasks by setting out the steps needed to complete them	-0.27	-1.57	-0.35	1.92
IS2G29E	To use tools to make diagrams that help solve problems	0.01	-1.52	-0.30	1.82
IS2G29F	To use simulations to help understand or solve real world problems	0.57	-1.42	-0.39	1.81
IS2G29G	To make flow diagrams to show the different parts of a process	0.68	-1.30	-0.32	1.62
IS2G29H	To record and evaluate data to understand and solve a problem	0.01	-1.60	-0.31	1.91
IS2G29I	To use real-world data to review and revise solutions to problems	0.00	-1.45	-0.35	1.80

Students' self-efficacy

Question 27 of the ICILS 2018 student questionnaire asked respondents to indicate how well they could do a series of tasks when using ICT (response categories were "I know how to do this," "I have never done this but I could work out how to do this," and "I do not think I could do this"). Twelve of the thirteen items from this question provided data for deriving two scales: *students' self-efficacy regarding the use of general applications* (S_GENEFF) and *students' self-efficacy regarding the use of specialist applications* (S_SPECEFF).

Figure 12.9 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with items from the scale. The model fit was very good, and there was a moderate correlation between the two latent factors (0.49). When reviewing measurement invariance using multiple-group models with different constraints, the results showed that model fit remained equally satisfactory with more constrained models suggesting a high degree of measurement invariance. S_GENEFF was derived from eight items and had an average reliability (Cronbach's alpha) of 0.83, with the coefficients ranging from 0.76 to 0.92 across participating countries (see Table 12.15). S_SPECEFF was based on four items and had an average scale reliability of 0.74, with coefficients ranging from 0.68 to 0.79. The higher values on these two scales represent a greater degree of self-efficacy. The item parameters for the two scales used for scaling are recorded in Table 12.16.

Figure 12.9: Confirmatory factor analysis of items measuring students' ICT self-efficacy



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.075	0.082	0.079	0.082
CFI	0.94	0.94	0.935	0.92
TLI	0.92	0.93	0.931	0.93

Table 12.15: Reliabilities for scales measuring students' ICT self-efficacy

Country	Scale reliability (Cronbach's alpha)	
	S_GENEFF	S_SPECEFF
Chile	0.81	0.73
Denmark	0.76	0.75
Finland	0.82	0.73
France	0.80	0.68
Germany	0.83	0.75
Italy	0.77	0.72
Kazakhstan	0.92	0.74
Korea, Republic of	0.89	0.79
Luxembourg	0.85	0.74
Moscow (Russian Federation)	0.89	0.71
North Rhine-Westphalia (Germany)	0.81	0.72
Portugal	0.81	0.73
United States	0.84	0.76
Uruguay	0.83	0.73
ICILS 2018 average	0.83	0.73

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.16: Item parameters for scales measuring students' ICT self-efficacy

Scale or item	Question/item wording	Item parameters		
		Delta	Tau(1)	Tau(2)
S_GENEFF	*How well can you do each of these tasks when using ICT?*			
IS2G27A	Edit digital photographs or other graphic images	0.17	-0.57	0.57
IS2G27C	Write or edit text for a school assignment	-0.28	-0.16	0.16
IS2G27D	Search for and find relevant information for a school project on the Internet	-0.41	-0.16	0.16
IS2G27I	Create a multi-media presentation (with sound, pictures, or video)	0.65	-0.78	0.78
IS2G27J	Upload text, images, or video to an online profile	0.06	-0.44	0.44
IS2G27K	Insert an image into a document or message	-0.29	-0.12	0.12
IS2G27L	Install a program or [app]	-0.20	-0.07	0.07
IS2G27M	Judge whether you can trust information you find on the Internet	0.30	-0.71	0.71
S_SPECEFF	*How well can you do each of these tasks when using ICT?*			
IS2G27B	Create a database (e.g. using [Microsoft Access ®])	-0.06	-1.15	1.15
IS2G27E	Build or edit a webpage	-0.28	-1.17	1.17
IS2G27G	Create a computer program, macro, or [app] e.g. in [Basic, Visual Basic])	0.62	-1.04	1.04
IS2G27H	Set up a local area network of computers or other ICT	-0.28	-0.69	0.69

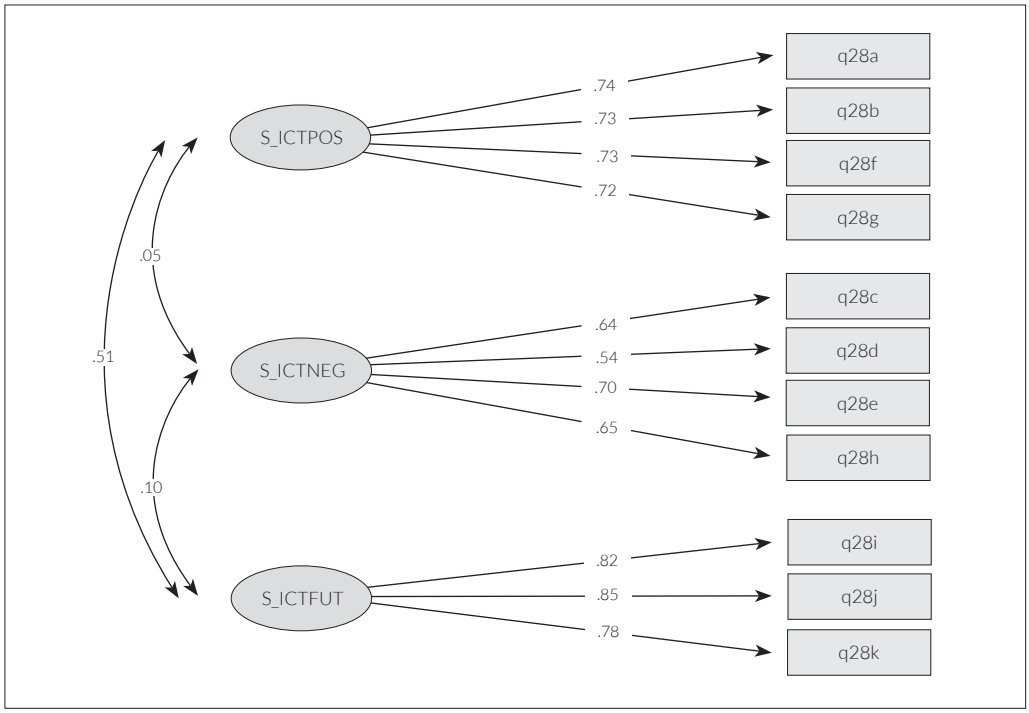
Students' perceptions of ICT

In Question 28 of the student questionnaire, respondents were asked provide their level of agreement ("strongly agree," "agree," "disagree," "strongly disagree") to a series of statements about ICT. The following scales were derived from the 11 items in the question:

- *Students' perceptions of positive outcomes of ICT for society* (S_ICTPOS)
- *Students' perceptions of negative outcomes of ICT for society* (S_ICTNEG)
- *Students' expectations of future ICT use for work and study* (S_ICTFUT)

Higher scores on these scales corresponded to stronger views (whether they be positive views, negative views, or greater expectations). A confirmatory factor analysis of the items in the question was run (see Figure 12.10), and supported a three-dimensional model. The model had a good fit, even when constraints were placed, which broadly suggests a relatively high level of measurement invariance. There was a strong positive correlation between S_ICTPOS and S_ICTFUT latent factors, but S_ICTNEG had only weak correlations with the other two scales.

Figure 12.10: Confirmatory factor analysis of items measuring students' perceptions of ICT



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.042	0.063	0.064	0.075
CFI	0.98	0.97	0.96	0.92
TLI	0.98	0.96	0.95	0.94

The average reliability (Cronbach's alpha) across countries for S_ICTPOS was 0.75 (ranging from 0.76 to 0.85), for S_ICTNEG it was 0.66 (ranging from 0.62 to 0.72), and for S_ICTFUT it was 0.80 (ranging from 0.74 to 0.85) (see Table 12.17). Item parameters for the three scales are presented in Table 12.18.

Table 12.17: Reliabilities for scales measuring students' perceptions of ICT

Country	Scale reliability (Cronbach's alpha)		
	S_ICTPOS	S_ICTNEG	S_ICTFUT
Chile	0.75	0.65	0.80
Denmark	0.67	0.63	0.84
Finland	0.81	0.67	0.83
France	0.73	0.68	0.82
Germany	0.72	0.65	0.82
Italy	0.73	0.63	0.74
Kazakhstan	0.79	0.68	0.79
Korea, Republic of	0.85	0.72	0.84
Luxembourg	0.73	0.64	0.80
Moscow (Russian Federation)	0.78	0.68	0.80
North Rhine-Westphalia (Germany)	0.70	0.62	0.85
Portugal	0.74	0.66	0.75
United States	0.75	0.67	0.81
Uruguay	0.77	0.68	0.75
ICILS 2018 average	0.75	0.66	0.81

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.18: Item parameters for scales measuring students' perceptions of ICT

Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
S_ICTPOS	*How much do you agree or disagree with the following statements about ICT?*				
IS2G28A	Advances in ICT usually improve people's living conditions.	0.01	-1.95	-0.79	2.75
IS2G28B	ICT helps us to understand the world better.	-0.11	-2.03	-0.70	2.73
IS2G28F	ICT is valuable to society.	0.02	-1.98	-0.64	2.62
IS2G28G	Advances in ICT bring many social benefits.	0.08	-2.06	-0.58	2.64
S_ICTNEG	*How much do you agree or disagree with the following statements about ICT?*				
IS2G28C	Using ICT makes people more isolated in society.	0.00	-1.65	0.03	1.62
IS2G28D	With more ICT there will be fewer jobs.	0.47	-1.62	0.22	1.40
IS2G28E	People spend far too much time using ICT.	-0.48	-1.18	-0.27	1.45
IS2G28H	Using ICT may be dangerous for people's health.	0.02	-1.33	-0.27	1.59
S_ICTFUT	*How much do you agree or disagree with the following statements about ICT?*				
IS2G28I	I would like to study subjects related to ICT after [secondary school]	0.34	-1.90	-0.07	1.97
IS2G28J	I hope to find a job that involves advanced ICT	0.24	-1.96	-0.06	2.02
IS2G28K	Learning how to use ICT applications will help me to do the work I am interested in	-0.57	-1.79	-0.40	2.19

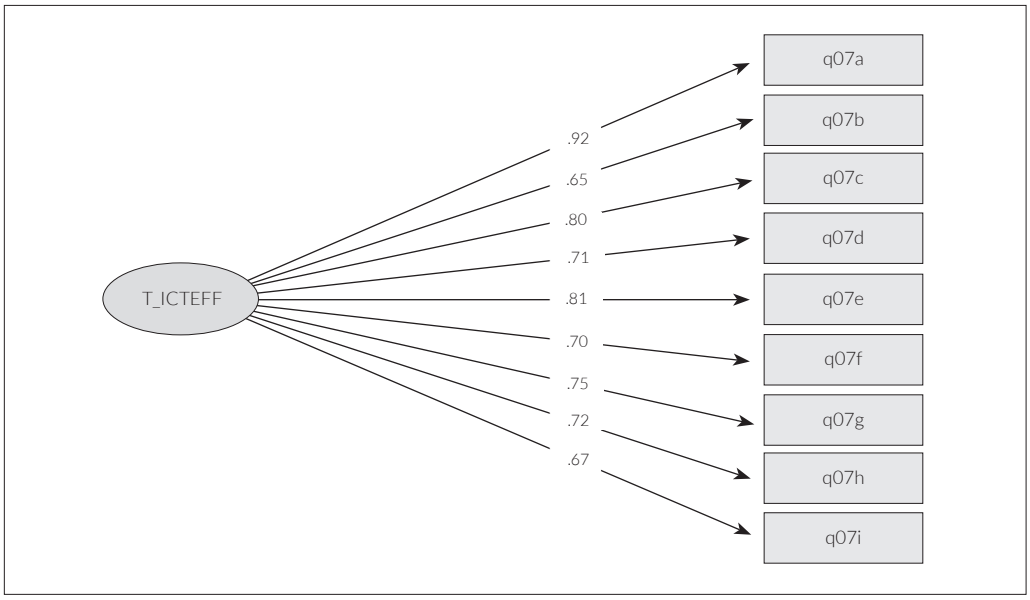
Teacher questionnaire

Teachers' ICT self-efficacy

Question 7 of the ICILS 2018 teacher questionnaire asked respondents to indicate how well they could do nine different school-related tasks using ICT. They were asked to select "I know how to do this," "I haven't done this but I could find out how," or "I do **not** think I could do this." All items were used to derive a scale on *teachers' ICT self-efficacy* (T_ICTEFF). Higher scale scores corresponded to a greater degree of teacher self-efficacy in using ICT for these tasks.

A confirmatory factor analysis (see Figure 12.11) reveals the model was highly satisfactory. When comparing the configural and scalar multiple-group model,[11] the model is somewhat less satisfactory suggesting some variation in measurement properties for this model. The items in the scale were shown to be highly reliable. The average reliability (Cronbach's alpha) across countries was 0.81 (ranging from 0.73 to 0.82) (see Table 12.19). The item parameters for the scale are presented in Table 12.20.

Figure 12.11: Confirmatory factor analysis of items measuring teachers' ICT self-efficacy



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.044	0.041	N/A	0.071
CFI	0.98	0.99	N/A	0.94
TLI	0.97	0.98	N/A	0.95

11 Please note that there was no convergence for the metric model.

Table 12.19: Reliabilities for scale measuring teachers' ICT self-efficacy

Country	Scale reliability (Cronbach's alpha)
	T_ICTEFF
Chile	0.74
Denmark	0.73
Finland	0.79
France	0.80
Germany	0.80
Italy	0.82
Kazakhstan	0.90
Korea, Republic of	0.84
Luxembourg	0.78
Moscow (Russian Federation)	0.77
North Rhine-Westphalia (Germany)	0.80
Portugal	0.79
United States	0.88
Uruguay	0.82
ICILS 2018 average	0.80

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.20: Item parameters for scale measuring teachers' ICT self-efficacy

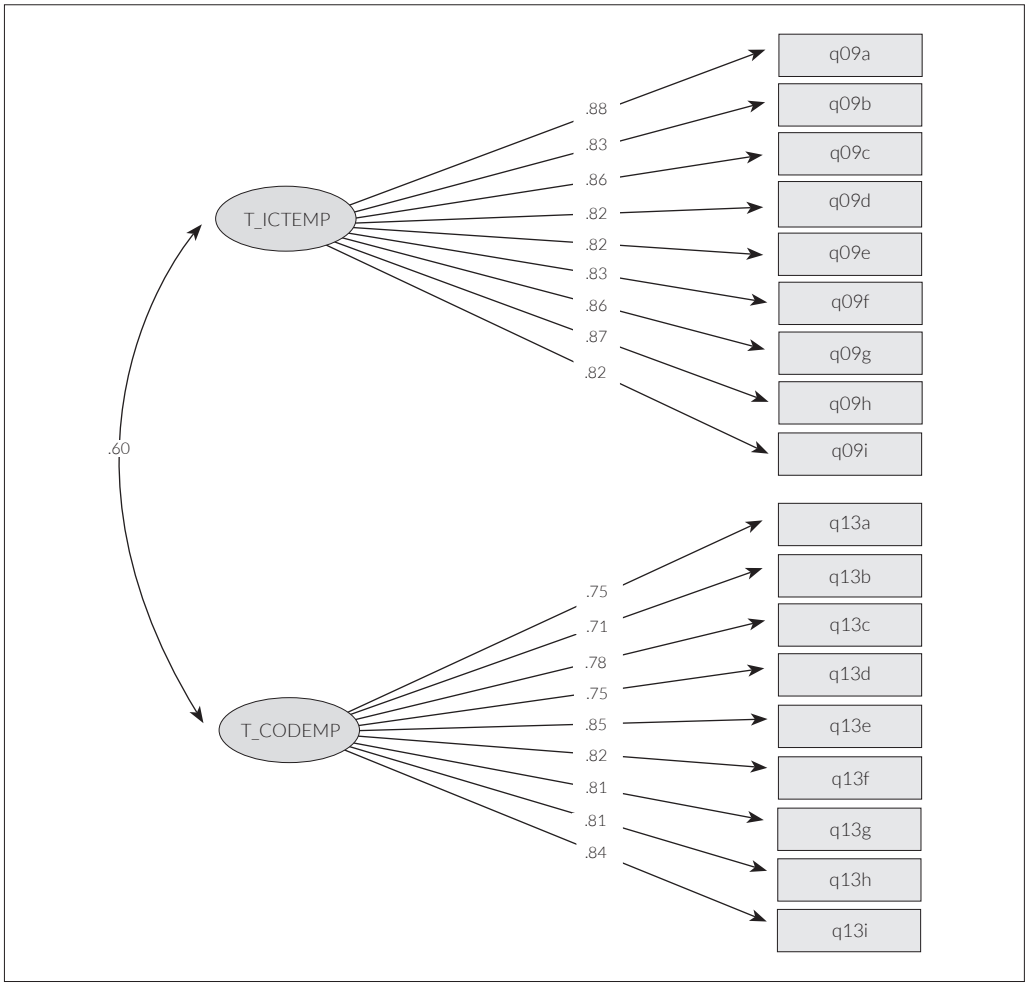
Scale or item	Question/item wording	Item parameters		
		Delta	Tau(1)	Tau(2)
T_ICTEFF	*How well can you do these tasks using ICT?*			
IT2G07A	Find useful teaching resources on the Internet	-1.18	0.70	-0.70
IT2G07B	Contribute to a discussion forum/user group on the Internet (eg. a wiki or blog)	0.58	-1.48	1.48
IT2G07C	Produce presentations (e.g. [PowerPoint® or a similar program]), with simple animation functions	-0.15	-0.22	0.22
IT2G07D	Use the Internet for online purchases and payments	-0.46	-0.36	0.36
IT2G07E	Prepare lessons that involve the use of ICT by students	-0.52	-0.68	0.68
IT2G07F	Using a spreadsheet program (e.g. [Microsoft Excel ®]) for keeping records or analyzing data	0.51	-0.79	0.79
IT2G07G	Assess student learning	-0.27	-0.94	0.94
IT2G07H	Collaborate with others using shared resources such as [Google Docs®], [Padlet]	0.80	-1.24	1.24
IT2G07I	Use a learning management system (e.g. [Moodle], [Blackboard], [Edmodo])	0.69	-1.27	1.27

Teachers' emphasis on developing ICT skills and coding skills

Teachers were asked in Question 9 to indicate the emphasis they had given to developing different ICT-based capabilities to students in the reference class (selecting from "strong emphasis," "some emphasis," "little emphasis," or "no emphasis"). The nine items in the question were used to derive a scale of *teachers' emphasis on developing ICT capabilities in class* (T_ICTEMP).

Similarly, Question 13 required teachers to indicate the emphasis they had given to teaching different skills related to coding in the reference class (using the same response options as in Question 9). The nine items in this question were used to derive a scale of *teacher emphasis of teaching CT-related tasks* (T_CODEMP). Higher scores on either scale corresponds to greater emphasis on developing ICT-based capabilities and coding skills.

Figure 12.12: Confirmatory factor analysis of items measuring teachers' emphasis on learning of ICT and coding tasks



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.066	0.063	N/A	0.075
CFI	0.96	0.96	N/A	0.93
TLI	0.95	0.96	N/A	0.94

Figure 12.12 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with items from the two scales. The model fit was highly satisfactory. When reviewing measurement invariance using multiple-group models with different constraints, the model fit changed only marginally which indicates a relatively high degree of invariance for this model. There was a high correlation between the two latent factors in the model (0.60). Both scales were highly reliable (see Table 12.21). The average reliability (Cronbach's alpha) across countries for both was 0.90 (ranging from 0.86 to 0.94 for T_ICTEMP, and between 0.88 and 0.93 for T_CODEMP). The item parameters for both scales are presented in Table 12.22

Table 12.21: Reliabilities for scales measuring teachers' emphasis on learning of ICT and coding tasks

Country	Scale reliability (Cronbach's alpha)	
	T_ICTEMP	T_CODEMP
Chile	0.90	0.92
Denmark	0.87	0.88
Finland	0.91	0.88
France	0.91	0.89
Germany	0.90	0.89
Italy	0.90	0.90
Kazakhstan	0.89	0.91
Korea, Republic of	0.92	0.93
Luxembourg	0.93	0.92
Moscow (Russian Federation)	0.86	0.91
North Rhine-Westphalia (Germany)	0.90	0.89
Portugal	0.92	0.91
United States	0.94	0.92
Uruguay	0.88	0.91
ICILS 2018 average	0.90	0.90

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.22: Item parameters for scales measuring teachers' emphasis on learning of ICT and coding tasks

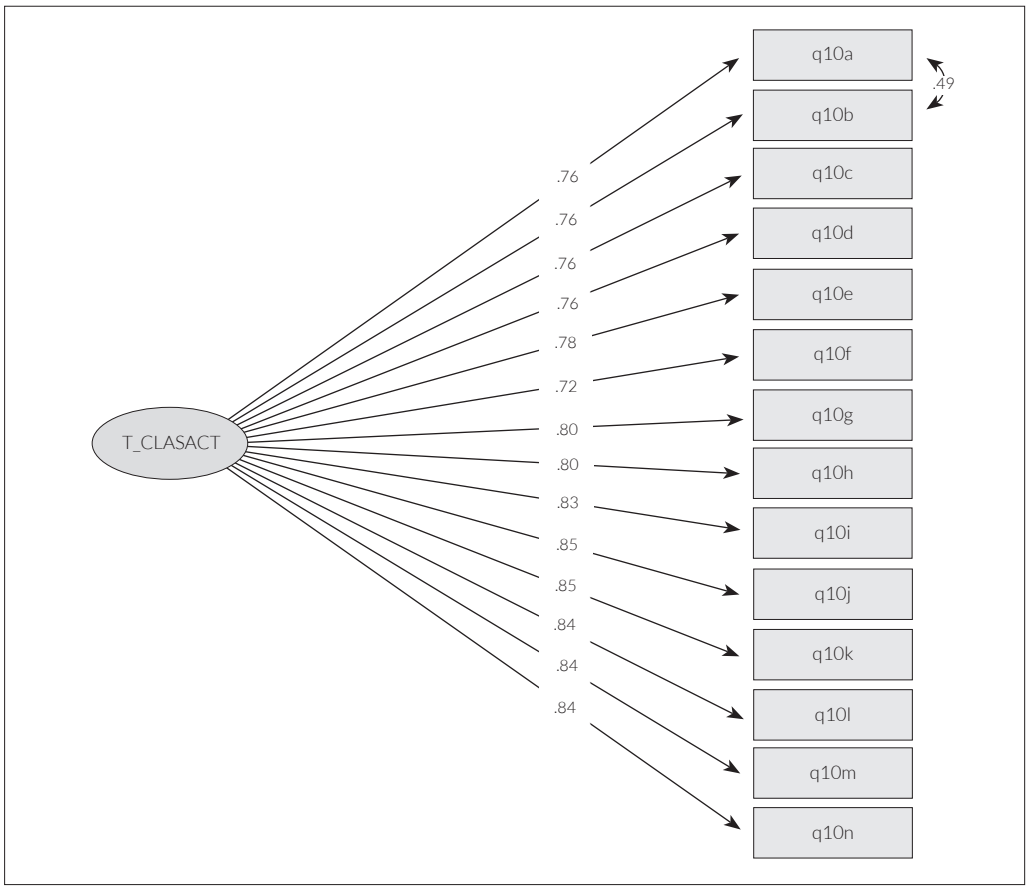
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
T_ICTEMP	*In your teaching the reference class in this school year, how much emphasis have you given to developing the following ICT-based capabilities in your students?*				
IT2G09A	To access information efficiently	-0.84	-2.12	-0.49	2.61
IT2G09B	To display information for a given audience/purpose	-0.37	-2.00	-0.47	2.46
IT2G09C	To evaluate the credibility of digital information	-0.20	-1.95	-0.25	2.20
IT2G09D	To share digital information with others	0.09	-2.02	-0.33	2.35
IT2G09E	To use computer software to construct digital work products (e.g. presentations, documents, images and diagrams)	-0.29	-1.61	-0.30	1.92
IT2G09F	To provide digital feedback on the work of others (such as classmates)	1.23	-1.87	-0.22	2.09
IT2G09G	To explore a range of digital resources when searching for information	-0.19	-1.87	-0.35	2.22
IT2G09H	To provide references for digital information sources	0.31	-1.70	-0.34	2.04
IT2G09I	To understand the consequences of making information publically available online	0.26	-1.57	-0.19	1.76
T_CODEMP	*In your teaching of the reference class this school year, how much emphasis have you given to teaching the following skills?*				
IT2G13A	To display information in different ways	-1.26	-1.82	-0.70	2.52
IT2G13B	To break a complex process into smaller parts	-0.74	-1.70	-0.61	2.30
IT2G13C	To understand diagrams that describe or show real-world problems	-0.06	-1.48	-0.51	2.00
IT2G13D	To plan tasks by setting out the steps needed to complete them	-0.75	-1.72	-0.48	2.20
IT2G13E	To use tools making diagrams that help solve problems	0.71	-1.48	-0.41	1.89
IT2G13F	To use simulations to help understand or solve real-world problems	0.72	-1.41	-0.43	1.84
IT2G13G	To make flow diagrams to show the different parts of a process	1.31	-1.38	-0.39	1.77
IT2G13H	To record and evaluate data to understand and solve a problem	0.11	-1.46	-0.54	1.99
IT2G13I	To use real-world data to review and revise solutions to problems	-0.04	-1.43	-0.54	1.97

Teachers' use of ICT for class activities

Question 10 of the ICILS 2018 teacher questionnaire asked teachers to select how often students in their reference class used ICT for different activities (they could choose from "they do not engage in this activity," "they never use ICT in this activity," "they sometimes use ICT in this activity," "they often use ICT in this activity," or "they always use ICT in this activity"). The 14 items in the question were used to derive a scale of *teachers' use of ICT for classroom activities* (T_CLASACT), where responses from teachers who indicated that their students did not engage in an activity were treated as missing values. Higher scale scores corresponded to more frequent use of ICT for these activities.

Figure 12.13 illustrates the results of the confirmatory factor analysis assuming a one-dimensional model with items from the scale. The analysis showed only marginally satisfactory model fit once residual variance for two items with similar content was taken into account (for items a and b). The model was equally marginally satisfactory for multiple-group models with different constraints, which suggests a relative high level of measurement invariance. The average reliability (Cronbach's alpha) of the scale was 0.94 (ranging from 0.92 to 0.96) (see Table 12.23). Table 12.24 shows the item parameters for the scale that was used to derive the IRT scale scores.

Figure 12.13: Confirmatory factor analysis of items measuring teachers' use of ICT for class activities



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.091	0.097	0.095	0.095
CFI	0.96	0.96	0.95	0.94
TLI	0.96	0.95	0.95	0.95

Table 12.23: Reliabilities for scale measuring teachers' use of ICT for class activities

Country	Scale reliability (Cronbach's alpha)
	T_CLASACT
Chile	0.94
Denmark	0.92
Finland	0.92
France	0.94
Germany	0.94
Italy	0.92
Kazakhstan	0.94
Korea, Republic of	0.95
Luxembourg	0.94
Moscow (Russian Federation)	0.92
North Rhine-Westphalia (Germany)	0.92
Portugal	0.96
United States	0.95
Uruguay	0.95
ICILS 2018 average	0.93

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.24: Item parameters for scale measuring teachers' use of ICT for class activities

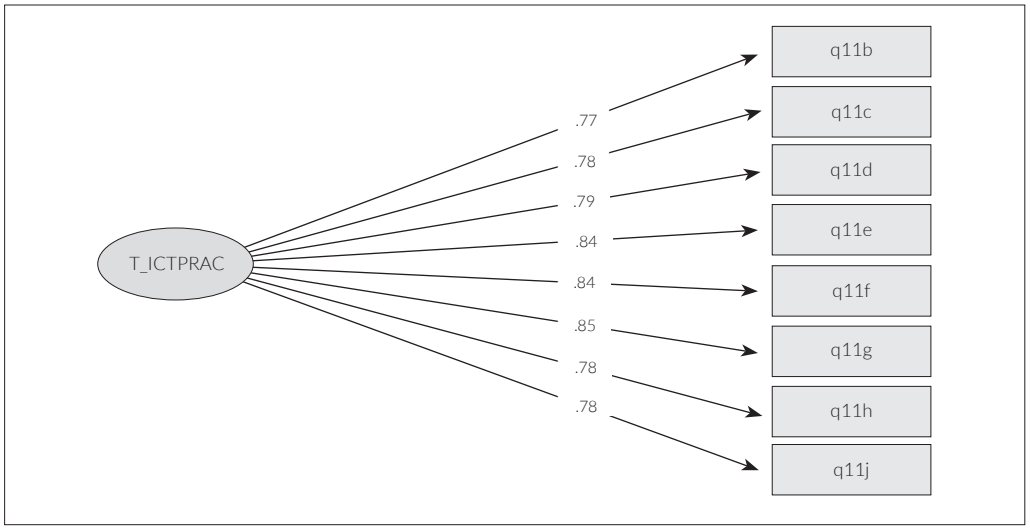
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
T_CLASACT	*How often do students in your reference class use ICT for the following activities?*				
IT2G10A	Work on extended projects (i.e. lasting over a week)	-0.73	-2.73	0.86	1.87
IT2G10B	Work on short assignments (i.e. within one week)	-0.67	-2.91	0.56	2.35
IT2G10C	Explain and discuss ideas with other students	0.73	-2.54	0.31	2.23
IT2G10D	Submit completed work for assessment	-0.32	-2.28	0.49	1.79
IT2G10E	Work individually on learning materials at their own pace	-0.07	-2.67	0.42	2.25
IT2G10F	Undertake open-ended investigations or field work	0.03	-2.69	0.59	2.10
IT2G10G	Reflect on their learning experiences (e.g. by using a learning log)	0.99	-1.92	0.36	1.57
IT2G10H	Communicate with students in other schools on projects	0.63	-2.19	0.27	1.92
IT2G10I	Plan a sequence of learning activities for themselves	0.97	-2.02	0.32	1.70
IT2G10J	Analyze data	0.22	-2.48	0.54	1.94
IT2G10K	Evaluate information resulting from a search	0.11	-2.49	0.53	1.97
IT2G10L	Collect data for a project	-0.90	-2.70	0.67	2.02
IT2G10M	Create visual products or videos	-0.76	-2.53	0.74	1.79
IT2G10N	Share products with other students	-0.22	-2.32	0.50	1.82

Teachers' use of ICT for teaching practices

In Question 11, respondents were asked to indicate how often they use ICT for 10 different practices related to teaching of the reference class (response options were "I do not use this practice with the reference class," "I never use ICT with this practice," "I sometimes use ICT with this practice," "I often use ICT with this practice," and "I always use ICT with this practice"). Eight of the 10 items were used to derive the scale *teachers' use of ICT for teaching practices in class* (T_ICTPRAC) where responses from teachers who indicated that they did not use this practice with the reference class were treated as missing values. Higher scale scores corresponded to more frequent use of ICT with the different practices.

Figure 12.14 illustrates the results of the confirmatory factor analysis assuming a one-dimensional model with items from the scale. The model fit was satisfactory for the pooled data set. When reviewing measurement invariance using multiple-group models with different constraints, the model fit changed only marginally which indicates a relatively high degree of invariance for this model. The average reliability (Cronbach's alpha) of the scale across countries was high (0.90), ranging from 0.86 to 0.93 (see Table 12.25). The item parameters used to derive the IRT scale scores are presented in Table 12.26.

Figure 12.14: Confirmatory factor analysis of items measuring teachers' use of ICT for teaching practices



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.063	0.074	0.076	0.080
CFI	0.99	0.99	0.98	0.97
TLI	0.99	0.98	0.98	0.98

Table 12.25: Reliabilities for scale measuring teachers' use of ICT for teaching practices

Country	Scale reliability (Cronbach's alpha)
	T_ICTPRAC
Chile	0.92
Denmark	0.88
Finland	0.86
France	0.89
Germany	0.88
Italy	0.87
Kazakhstan	0.92
Korea, Republic of	0.92
Luxembourg	0.89
Moscow (Russian Federation)	0.90
North Rhine-Westphalia (Germany)	0.86
Portugal	0.91
United States	0.91
Uruguay	0.93
ICILS 2018 average	0.89

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.26: Item parameters for scale measuring teachers' use of ICT for teaching practices

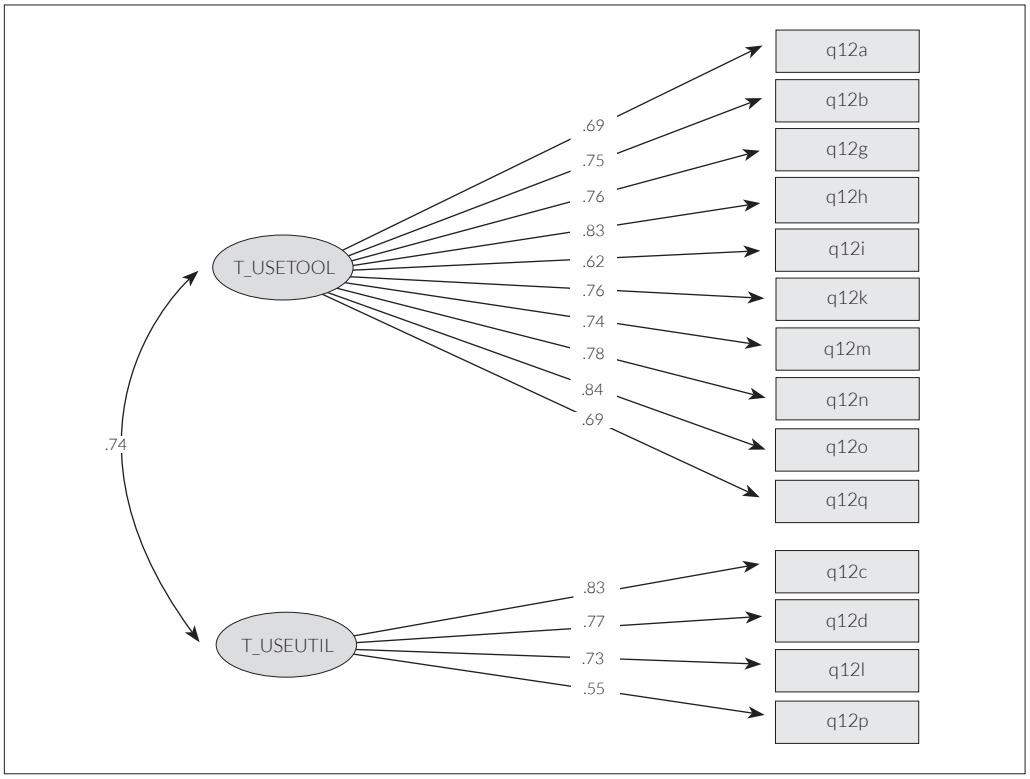
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
T_ICTPRAC	*How often do you use ICT in the following practices when teaching your reference class?*				
IT2G11B	The provision of remedial or enrichment support to individual students or small groups of students	-0.20	-2.41	0.26	2.15
IT2G11C	The support of student-led whole-class discussions and presentations	-0.34	-2.34	0.21	2.12
IT2G11D	The assessment of students' learning through tests	-0.10	-1.72	0.20	1.52
IT2G11E	The provision of feedback to students on their work	0.21	-2.07	0.27	1.80
IT2G11F	The reinforcement of learning of skills through repetition of examples	-0.28	-2.37	0.25	2.12
IT2G11G	The support of collaboration among students	0.25	-2.28	0.23	2.05
IT2G11H	The mediation of communication between students and experts or external mentors	0.77	-1.78	0.05	1.73
IT2G11J	The support of inquiry learning	-0.31	-2.35	0.37	1.98

Teachers' use of ICT tools in class

Question 12 of the ICILS 2018 teacher questionnaire asked teachers how often they used different ICT-related tools in the teaching of their reference classes. For each tool, respondents were asked to select either "never," "in some lessons," "in most lessons," or "in every or almost every lesson." Two scales were derived from the set of items: *teachers' use of digital learning tools* (T_USETOOL) and *teachers' use of general utility software* (T_USEUTIL). Higher scores on either scale reflects more frequent use of these types of tools.

Figure 12.15 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with items from the two scales. The model fit was satisfactory for the pooled dataset, however, with increasing constraints across different multiple-group models the fit became unsatisfactory, a finding which suggests a certain degree of variation in measurement characteristics. The correlation between the two latent factors was quite high (0.74). The average reliabilities (Cronbach's alpha) of the two scales across countries were 0.82 for T_USETOOL and 0.73 for T_USEUTIL (ranging from 0.71 to 0.91 for the former, and ranging from 0.65 to 0.82 for the latter) (see Table 12.27). Table 12.28 shows the item parameters for each of the two scales that were used to derive the IRT scale scores.

Figure 12.15: Confirmatory factor analysis of items measuring teachers' use of ICT tools in class



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.063	0.087	0.109	0.105
CFI	0.96	0.94	0.88	0.87
TLI	0.95	0.92	0.88	0.89

Table 12.27: Reliabilities for scales measuring teachers' use of ICT tools in class

Country	Scale reliability (Cronbach's alpha)	
	T_USETOOL	T_USEUTIL
Chile	0.87	0.81
Denmark	0.71	0.66
Finland	0.75	0.68
France	0.79	0.65
Germany	0.83	0.71
Italy	0.79	0.78
Kazakhstan	0.91	0.80
Korea, Republic of	0.90	0.78
Luxembourg	0.78	0.70
Moscow (Russian Federation)	0.89	0.82
North Rhine-Westphalia (Germany)	0.79	0.74
Portugal	0.80	0.76
United States	0.86	0.72
Uruguay	0.86	0.78
ICILS 2018 average	0.82	0.74

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.28: Item parameters for scales measuring teachers' use of ICT use in class

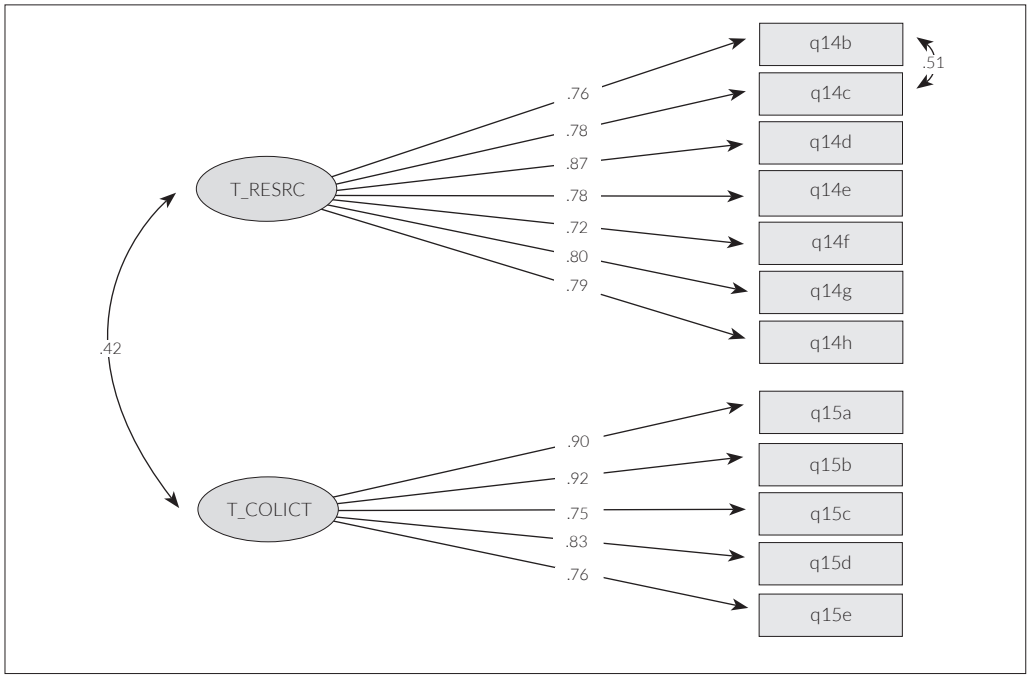
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
T_USETOOL	*How often did you use the following tools in your teaching of the reference class this school year?*				
IT2G12A	Practice programs or apps where you ask students questions (e.g. [Quizlet, Kahoot], [mathfessor])	0.02	-1.87	0.61	1.27
IT2G12B	Digital learning games	-0.16	-2.08	0.69	1.38
IT2G12G	Concept mapping software (e.g. [Inspiration ®], [Webspiration ®])	0.54	-1.18	0.26	0.93
IT2G12H	Simulations and modelling software (e.g. [NetLogo])	0.88	-0.72	-0.02	0.74
IT2G12I	A learning management system (e.g. [Edmodo], [Blackboard])	-1.03	-0.43	0.24	0.20
IT2G12K	Collaborative software (e.g. [Google Docs ®], [Onenote]) [Padlet])	-0.37	-1.25	0.29	0.96
IT2G12M	Interactive digital learning resources (e.g. learning objects)	-0.64	-1.35	0.24	1.11
IT2G12N	Graphing or drawing software	0.26	-1.15	0.18	0.97
IT2G12O	e-portfolios (e.g. [VoiceThread])	0.58	-0.57	-0.09	0.66
IT2G12Q	Social media (e.g. [Facebook, Twitter])	-0.08	-1.02	0.34	0.68
T_USEUTIL	*How often did you use the following tools in your teaching of the reference class this school year?*				
IT2G12C	Word-processor software (e.g. [Microsoft Word ®])	-0.24	-1.93	0.52	1.41
IT2G12D	Presentation software (e.g. [Microsoft PowerPoint ®])	-0.25	-2.05	0.55	1.50
IT2G12L	Computer-based information resources (e.g. topic-related websites, wikis, encyclopaedia)	0.10	-2.25	0.53	1.72
IT2G12P	Digital contents linked with textbooks	0.39	-1.37	0.39	0.98

Teachers' perceptions of ICT resources and teacher collaboration

Question 14 of the ICILS 2018 teacher questionnaire asked teachers to provide their level of agreement or disagreement ("strongly agree," "agree," "disagree," "strongly disagree") to a series of statements about ICT resources at their schools. Question 15 requested respondents to rate their agreement or disagreement ("strongly agree," "agree," "disagree," "strongly disagree") with statements about ICT-related teacher collaboration at school. Two scales were derived from the items included in these two questions: *teachers' perceptions of the availability of computer resources at school* (T_RESRC) and *teachers' perceptions of the collaboration between teachers when using ICT* (T_COLICT). Higher scores on either scale reflects higher levels of agreement with the statements used for measurement.

Figure 12.16 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with items from the two scales. The model fit was marginally satisfactory for the pooled dataset after allowing residuals for two items to be correlated (item a and b in Question 14). Across different multiple-group models the fit remained marginally satisfactory with least good fit for the scalar model, which suggests a certain degree of variation in measurement characteristics. The correlation between the two latent factors was moderately positive (0.42). The average reliabilities (Cronbach's alpha) of the two scales across countries were 0.87 for T_RESRC and 0.85 for T_COLICT (ranging from 0.83 to 0.93 for the former, and ranging from 0.77 to 0.92 for the latter) (see Table 12.29). Table 12.30 shows the item parameters for each of the two scales that were used to derive the IRT scale scores.

Figure 12.16: Confirmatory factor analysis of items measuring teachers' perceptions of ICT resources and teacher collaboration at school



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.086	0.084	0.081	0.092
CFI	0.96	0.97	0.97	0.95
TLI	0.95	0.97	0.97	0.96

Table 12.29: Reliabilities for scales measuring teachers' perceptions of ICT resources and teacher collaboration at school

Country	Scale reliability (Cronbach's alpha)	
	T_RESRC	T_COLICT
Chile	0.89	0.90
Denmark	0.84	0.83
Finland	0.83	0.86
France	0.85	0.86
Germany	0.89	0.77
Italy	0.87	0.89
Kazakhstan	0.93	0.85
Korea, Republic of	0.90	0.89
Luxembourg	0.84	0.85
Moscow (Russian Federation)	0.88	0.88
North Rhine-Westphalia (Germany)	0.87	0.81
Portugal	0.86	0.86
United States	0.90	0.92
Uruguay	0.88	0.89
ICILS 2018 average	0.87	0.85

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.30: Item parameters for scales measuring teachers' perceptions of ICT resources and teacher collaboration at school

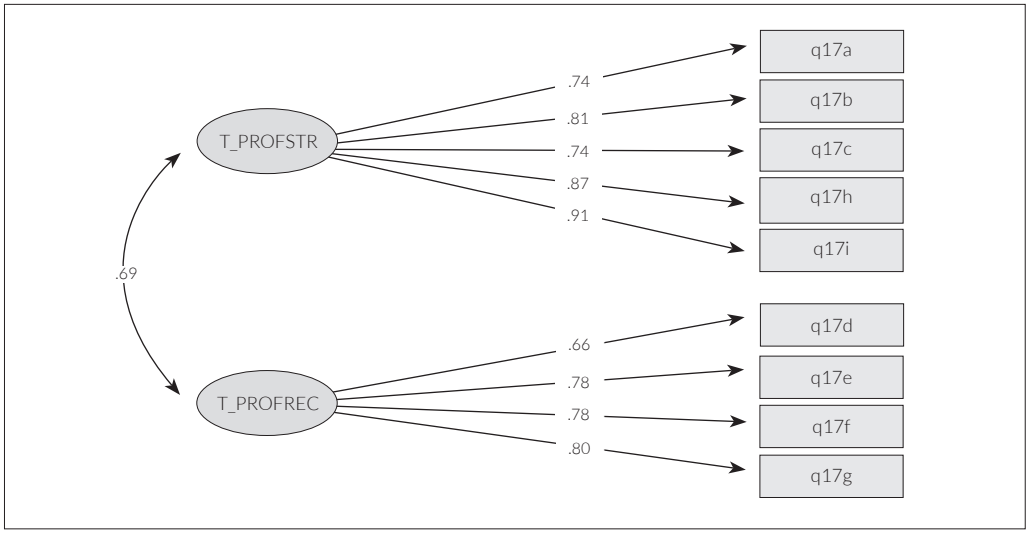
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
T_RESRC	*To what extent do you agree or disagree with the following statements about using ICT in teaching at your school?*				
IT2G14B	My school has sufficient ICT equipment (e.g. computers).	-0.42	-2.15	-0.12	2.26
IT2G14C	The computer equipment in our school is up-to-date.	-0.30	-2.29	-0.22	2.51
IT2G14D	My school has access to sufficient digital learning resources (e.g. learning software or [apps]).	-0.21	-2.50	-0.18	2.68
IT2G14E	My school has good connectivity (e.g. fast spped – same as in STable) to the Internet.	-0.06	-2.12	-0.30	2.42
IT2G14F	There is enough time to prepare lessons that incorporate ICT.	0.69	-2.72	-0.05	2.77
IT2G14G	There is sufficient opportunity for me to develop expertise in ICT.	0.21	-2.82	-0.23	3.05
IT2G14H	There is sufficient technical support to maintain ICT resources.	0.09	-2.44	-0.31	2.76
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?*				
IT2G15A	I work together with other teachers on improving the use of ICT in classroom teaching.	0.18	-3.62	-0.35	3.97
IT2G15B	I collaborate with colleagues to develop ICT-based lessons.	0.34	-3.87	-0.23	4.10
IT2G15C	I observe how other teachers use ICT in teaching.	0.01	-3.57	-0.69	4.26
IT2G15D	I discuss with other teachers how to use ICT in. teaching topics	-0.24	-3.67	-0.80	4.47
IT2G15E	I share ICT-based resources with other teachers in my school.	-0.29	-3.40	-0.62	4.03

Teachers' reports on ICT-related professional learning

Question 17 of the ICILS 2018 teacher questionnaire asked teachers to indicate their participation ("not at all," "once only," "more than once") in different ICT-related professional learning activities. We derived two scales from the set of items: *teacher participation in structured learning professional development related to ICT* (T_PROFSTR) and *teacher participation in reciprocal learning professional development related to ICT* (T_PROFREC). Higher scores on either scale reflects higher levels of participation in these two types of professional learning activities.

Figure 12.17 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with items from the two scales. The model fit was satisfactory for the pooled dataset, however, with more constraints across different multiple-group models the fit became only marginally satisfactory, a finding which suggests a certain degree of variation in measurement characteristics. The correlation between the two latent factors was quite high (0.69). The average reliabilities (Cronbach's alpha) of the two scales across countries were 0.76 for T_PROFSTR and 0.69 for T_PROFREC (ranging from 0.63 to 0.87 for the former, and ranging from 0.58 to 0.79 for the latter) (see Table 12.27). Table 12.28 shows the item parameters for each of the two scales that were used to derive the IRT scale scores.

Figure 12.17: Confirmatory factor analysis of items measuring teachers' participation in ICT-related professional learning



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.074	0.076	0.095	0.096
CFI	0.96	0.97	0.94	0.92
TLI	0.95	0.96	0.93	0.93

Table 12.27: Reliabilities for scales measuring teachers' participation in ICT-related professional learning

Country	Scale reliability (Cronbach's alpha)	
	T_PROFSTR	T_PROFREC
Chile	0.84	0.74
Denmark	0.75	0.64
Finland	0.63	0.67
France	0.72	0.69
Germany	0.72	0.62
Italy	0.79	0.68
Kazakhstan	0.87	0.79
Korea, Republic of	0.83	0.78
Luxembourg	0.73	0.73
Moscow (Russian Federation)	0.80	0.76
North Rhine-Westphalia (Germany)	0.64	0.58
Portugal	0.75	0.65
United States	0.82	0.77
Uruguay	0.80	0.74
ICILS 2018 average	0.76	0.69

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.28: Item parameters for scales measuring teachers' participation in ICT-related professional learning

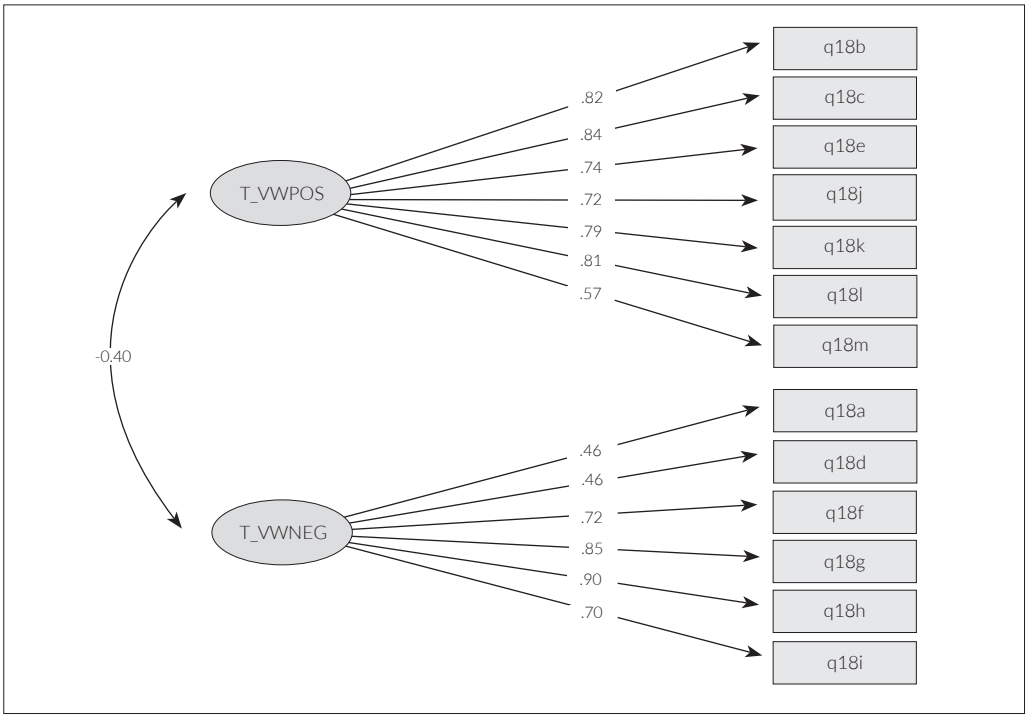
Scale or item	Question/item wording	Item parameters		
		Delta	Tau(1)	Tau(2)
T_PROFSTR	*How often have you participated in any of the following professional learning activities in the past two years?*			
IT2G17A	A course on ICT applications (e.g. word processing, presentations, internet use, spreadsheets, databases)	-0.62	-0.52	0.52
IT2G17B	A course or webinar on integrating ICT into teaching and learning	-0.34	-0.55	0.55
IT2G17C	Training on subject-specific digital teaching and learning resources	-0.45	-0.70	0.70
IT2G17H	A course on use of ICT for [students with special needs or specific learning difficulties]	0.84	-0.24	0.24
IT2G17I	A course on how to use ICT to support personalized learning by students	0.58	-0.31	0.31
T_PROFREC	*How often did you use the following tools in your teaching of the reference class this school year?*			
IT2G17D	Observations of other teachers using ICT in teaching	-0.40	0.10	-0.10
IT2G17E	An ICT-mediated discussion or forum on teaching and learning	0.39	0.19	-0.19
IT2G17F	The sharing of digital teaching and learning resources with others through a collaborative workspace	-0.31	0.08	-0.08
IT2G17G	Use of a collaborative workspace to jointly evaluate student work	0.33	0.31	-0.31

Teachers' perceptions of positive and negative outcomes of using ICT for teaching and learning

Question 18 of the ICILS 2018 teacher questionnaire asked teachers to provide their level of agreement or disagreement ("strongly agree," "agree," "disagree," "strongly disagree") to a series of statements about positive and negative outcomes of using ICT for teaching and learning. Two scales were derived from the set of items: *teachers' perceptions of positive outcomes when using ICT in teaching and learning* (T_VWPOS) and *teachers' perceptions of negative outcomes when using ICT in teaching and learning* (T_VWNEG). Higher scores on either scale reflects higher levels of agreement for each of these scales.

Figure 12.18 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with items from the two scales. The model fit was satisfactory for the pooled dataset, however, with increasing constraints across different multiple-group models the fit became unsatisfactory, a finding which suggests a certain degree of variation in measurement characteristics. There was a moderately high negative correlation between the two latent factors (-0.40).

Figure 12.18: Confirmatory factor analysis of items measuring teachers' perceptions of positive and negative outcomes of using ICT for teaching and learning



Model fit indices:	Pooled sample	Multiple-group models		
		Configural	Metric	Scalar
RMSEA	0.063	0.094	N/A	0.092
CFI	0.96	0.94	N/A	0.91
TLI	0.95	0.93	N/A	0.93

The average reliabilities (Cronbach's alpha) of the two scales across countries were 0.83 for T_VWPOS and 0.80 for T_VWNEG (ranging from 0.79 to 0.87 for the former, and ranging from 0.76 to 0.85 for the latter) (see Table 12.33). Table 12.34 shows the item parameters for each of the two scales that were used to derive the IRT scale scores.

Table 12.33: Reliabilities for scales measuring teachers' perceptions of positive and negative outcomes of using ICT for teaching and learning

Country	Scale reliability (Cronbach's alpha)	
	T_VWPOS	T_VWNEG
Chile	0.86	0.81
Denmark	0.81	0.77
Finland	0.79	0.76
France	0.82	0.80
Germany	0.82	0.80
Italy	0.86	0.85
Kazakhstan	0.87	0.77
Korea, Republic of	0.83	0.82
Luxembourg	0.80	0.80
Moscow (Russian Federation)	0.84	0.82
North Rhine-Westphalia (Germany)	0.83	0.81
Portugal	0.84	0.82
United States	0.87	0.81
Uruguay	0.87	0.83
ICILS 2018 average	0.83	0.80

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.34: Item parameters for scales measuring teachers' perceptions of positive and negative outcomes of using ICT for teaching and learning

Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
T_VWPOS	*To what extent do you agree or disagree with the following practices and principles in relation to the use of ICT in teaching and learning?*				
IT2G18B	Helps students develop greater interest in learning.	-0.67	-2.84	-1.01	3.85
IT2G18C	Helps students to work at a level appropriate to their learning needs.	-0.42	-3.51	-0.66	4.17
IT2G18E	Helps students develop problem solving skills.	0.29	-3.66	-0.63	4.29
IT2G18J	Enables students to collaborate more effectively.	0.27	-3.93	-0.49	4.42
IT2G18K	Helps students develop skills in planning and self-regulation of their work.	0.61	-3.88	-0.41	4.29
IT2G18L	Improves academic performance of students.	0.73	-3.89	-0.41	4.30
IT2G18M	Enables students to access better sources of information.	-0.82	-2.87	-1.06	3.93
T_VWNEG	*To what extent do you agree or disagree with the following practices and principles in relation to the use of ICT in teaching and learning?*				
IT2G18A	Impedes concept formation by students.	0.93	-2.64	0.82	1.82
IT2G18D	Results in students copying material from Internet sources.	-0.99	-3.05	-0.02	3.07
IT2G18F	Distracts students from learning.	0.33	-3.26	0.55	2.71
IT2G18G	Results in poorer written expression among students.	-0.25	-2.93	0.29	2.63
IT2G18H	Results in poorer calculation and estimation skills among students.	0.09	-3.28	0.57	2.71
IT2G18I	Limits the amount of personal communication among students.	-0.10	-3.04	0.50	2.55

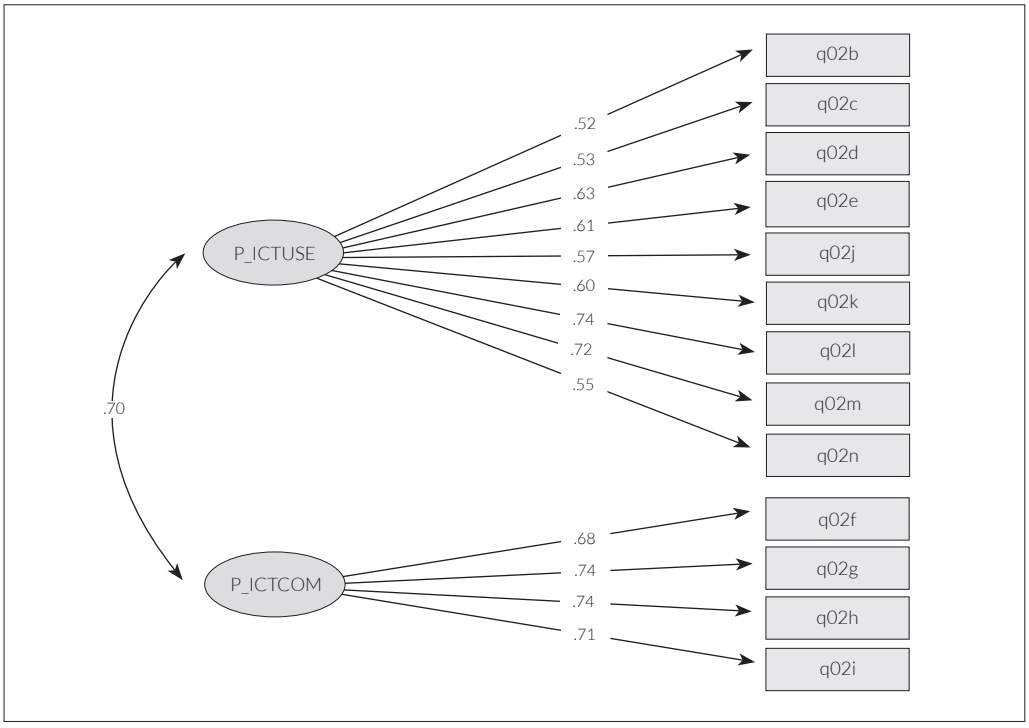
School questionnaires

School principals' use of ICT

The school principal questionnaire asked respondents to indicate how often they used ICT for different school related activities (for each item they could select from "never," "less than once a month," "at least once a month but not every week," "at least once a week but not every day," and "every day"). Nine of the items were used to derive the scale *principals' use of ICT for general school-related activities* (P_ICTUSE) and four of the remaining five items were used to derive the scale *principals' use of ICT for school-related communication activities* (P_ICTCOM). Higher scores on these two scales represent more frequent use of ICT.

A confirmatory factor analysis assuming a two-dimensional model with items from the two scales (see Figure 12.19) was satisfactory for the pooled ICILS 2018 sample. The two latent factors in the model were strongly correlated (0.70). Average reliabilities across countries for the two scales were satisfactory in most countries (see Table 12.35). P_ICTUSE had an average reliability of 0.79 (ranging from 0.68 to 0.87) and P_ICTCOM had an average reliability of 0.70 (ranging from 0.47 to 0.79). Table 12.36 records the IRT parameters for each of the two scales.

Figure 12.19: Confirmatory factor analysis of items measuring school principals' use of ICT



Model fit indices:	Pooled sample
RMSEA	0.058
CFI	0.92
TLI	0.90

Table 12.35: Reliabilities for scales measuring principals' use of ICT

Country	Scale reliability (Cronbach's alpha)	
	P_ICTUSE	P_ICTCOM
Chile	0.85	0.74
Denmark	0.77	0.54
Finland	0.73	0.78
France	0.73	0.76
Germany	0.81	0.67
Italy	0.81	0.66
Kazakhstan	0.80	0.81
Korea, Republic of	0.87	0.82
Luxembourg	0.78	0.47
Moscow (Russian Federation)	0.81	0.79
North Rhine-Westphalia (Germany)	0.68	0.62
Portugal	0.76	0.71
United States	0.83	0.83
Uruguay	0.78	0.59
ICILS 2018 average	0.78	0.70

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.36: Item parameters for scales measuring principals' use of ICT

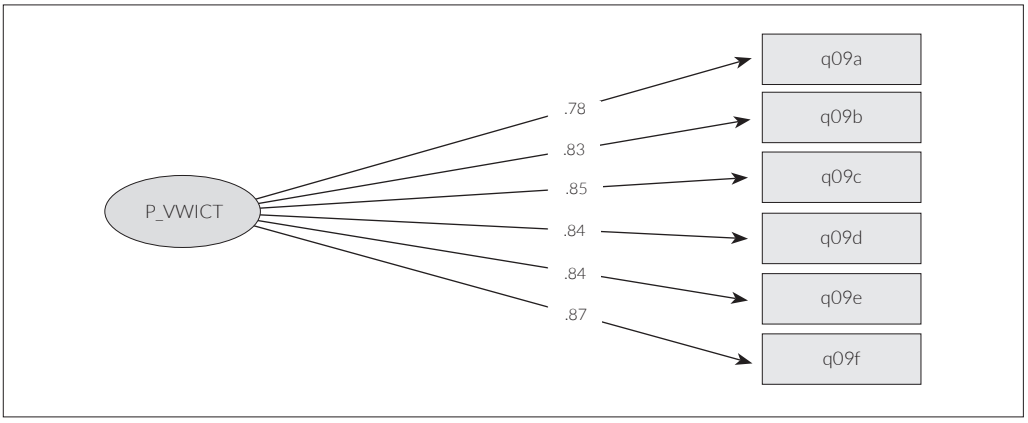
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
P_ICTUSE	*How often do you use ICT for the following activities?*				
IP2G02B	Provide information about an educational issue through a website	-0.15	-0.57	-0.66	1.24
IP2G02C	Look up records in a database (e.g. in a student information system)	-0.91	-0.01	-0.56	0.57
IP2G02D	Maintain, organize and analyze data (e.g. with a spreadsheet or database)	-0.55	-0.37	-0.43	0.80
IP2G02E	Prepare presentations	0.49	-1.31	-0.30	1.61
IP2G02J	Work with a learning management system (e.g. [Moodle])	0.56	0.13	-0.36	0.23
IP2G02K	Use social media to communicate with the wider community about school-related activities	0.45	0.37	-0.81	0.44
IP2G02L	Management of staff (e.g. scheduling, professional development)	-0.48	-0.46	-0.21	0.67
IP2G02M	Preparing the curriculum	0.47	-0.34	-0.32	0.66
IP2G02N	School financial management	0.11	-0.18	-0.35	0.53
P_ICTCOM	*How often do you use ICT for the following activities?*				
IP2G02F	Communicate with teachers in your school	-1.15	-0.03	-0.78	0.81
IP2G02G	Communicate with education authorities	-0.08	-1.19	-0.31	1.50
IP2G02H	Communicate with principals and senior staff in other schools	0.43	-1.14	-0.52	1.65
IP2G02I	Communicate with parents	0.80	-0.92	-0.59	1.51

School principals' views on using ICT

In Question 9 of the ICILS 2018 principal questionnaire, respondents were given seven different ICT-related outcomes of education in their school, and were asked to rate their perceived level of importance (selecting from "very important," "quite important," "somewhat important," and "not important"). The first six of the seven items in the question were used to derive the scale *principals' views on using ICT for educational outcomes* (P_VWICT). Higher scale scores corresponded to higher levels of importance assigned to ICT-related skills as an outcome of learning.

Figure 12.20 presents the results of the confirmatory factor analysis. The one-dimensional model had a highly satisfactory fit for the pooled ICILS 2018 dataset. Table 12.37 shows the scale reliabilities (Cronbach's alpha) across countries, which were quite high (on average 0.87, ranging from 0.75 to 0.96 across countries). Table 12.38 shows the IRT item parameters that were used to derive the final scale scores.

Figure 12.20: Confirmatory factor analysis of items measuring school principals' views on using ICT



Model fit indices:	Pooled sample
RMSEA	0.033
CFI	1.00
TLI	1.00

Table 12.37: Reliabilities for scale measuring principals' views on using ICT

Country	Scale reliability (Cronbach's alpha)
	P_VWICT
Chile	0.96
Denmark	0.81
Finland	0.84
France	0.89
Germany	0.84
Italy	0.86
Kazakhstan	0.85
Korea, Republic of	0.89
Luxembourg	0.82
Moscow (Russian Federation)	0.75
North Rhine-Westphalia (Germany)	0.87
Portugal	0.85
United States	0.93
Uruguay	0.92
ICILS 2018 average	0.85

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.38: Item parameters for scale measuring principals' views of using ICT

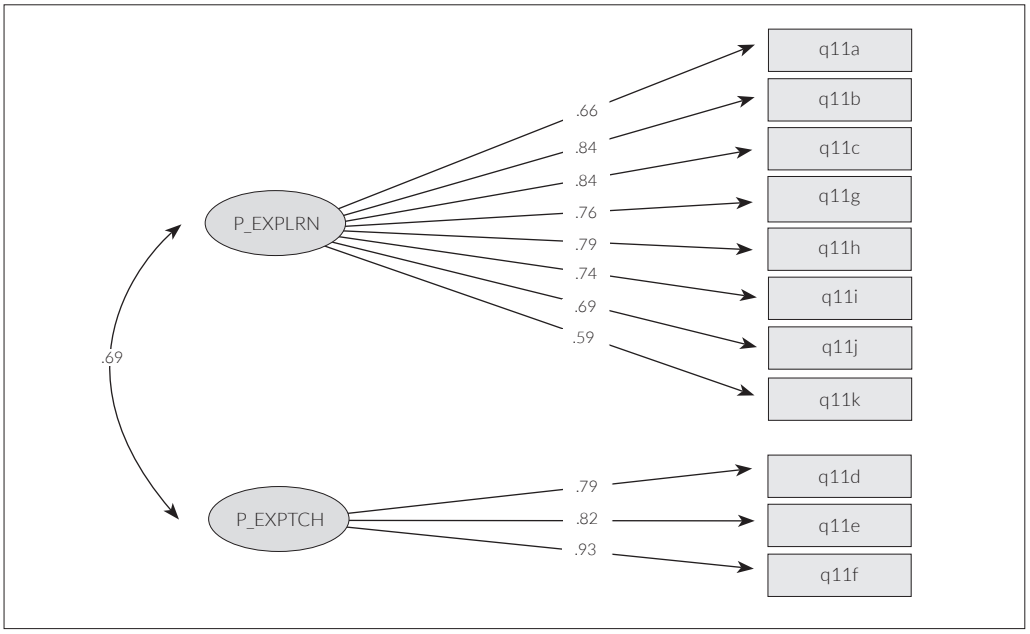
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
P_VWICT	*How important is each of the following outcomes of education in your school?*				
IP2G09A	The development of students' basic computer skills (e.g. internet use, email, word processing, presentation software)	-0.06	-2.61	-0.58	3.19
IP2G09B	The development of students' skills in using ICT for collaboration with others	0.39	-4.10	0.30	3.80
IP2G09C	The use of ICT for facilitating students' responsibility for their own learning	0.66	-3.83	0.21	3.62
IP2G09D	The use of ICT to augment and improve students' learning	0.07	-4.16	0.22	3.94
IP2G09E	The development of students' understanding and skills relating to safe and appropriate use of ICT	-0.67	-3.54	-0.01	3.55
IP2G09F	The development of students' proficiency in accessing and using information with ICT	-0.39	-3.74	-0.07	3.81

School principals' reports on expected ICT knowledge and skills of teachers

In Question 11, school principals were asked whether teachers in their school were expected to acquire knowledge and skills in a range of different activities related to ICT. For each activity they were asked to select either "expected and required," "expected but not required," or "not expected." Data from eight items were used to derive the scale *principals' reports on expectations of ICT use by teachers* (P_EXPLRN) and data from the remaining three items provided the basis for the scale *principals' reports on expectations for teacher collaboration using ICT* (P_EXPTCH). Higher scores on these two scales represent higher levels of expectations from the principals.

Figure 12.21 provides the results of the confirmatory factor analysis of data from the items in the question. Here we can see that the two-dimensional model had a marginally satisfactory fit for the pooled dataset. The results also showed a relative high positive correlation between the two latent factors (0.69). Table 12.39 displays the scale reliabilities (Cronbach's alpha) for the two scales. P_EXPLRN had an average reliability across countries of 0.78 (ranging from 0.66 to 0.89) whereas P_EXPTCH had an average reliability across countries of 0.70 (ranging from 0.59 to 0.86). Table 12.40 shows the IRT item parameters that we used to derive the final scale scores for each of the two scales.

Figure 12.21: Confirmatory factor analysis of items measuring principals' reports on expected ICT knowledge and skills of teachers



Model fit indices:	Pooled sample
RMSEA	0.083
CFI	0.95
TLI	0.93

Table 12.39: Reliabilities for scales measuring principals' reports on expected ICT knowledge and skills of teachers

Country	Scale reliability (Cronbach's alpha)	
	P_EXPLRN	P_EXPTCH
Chile	0.89	0.78
Denmark	0.73	0.60
Finland	0.74	0.63
France	0.74	0.76
Germany	0.77	0.64
Italy	0.75	0.65
Kazakhstan	0.78	0.73
Korea, Republic of	0.89	0.86
Luxembourg	0.81	0.86
Moscow (Russian Federation)	0.76	0.69
North Rhine-Westphalia (Germany)	0.66	0.72
Portugal	0.69	0.59
United States	0.81	0.65
Uruguay	0.81	0.65
ICILS 2018 average	0.77	0.71

Table 12.40: Item parameters for scales measuring principals' reports on expected ICT knowledge and skills of teachers

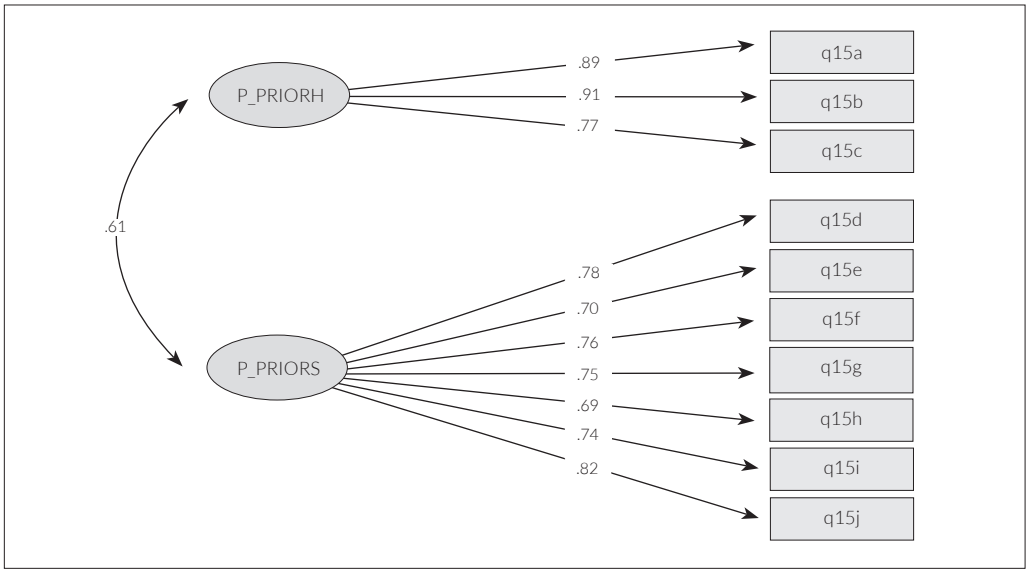
Scale or item	Question/item wording	Item parameters		
		Delta	Tau(1)	Tau(2)
P_EXPLRN	*Are teachers in your school expected to acquire knowledge and skills in each of the following activities?*			
IP2G11A	Integrate Web-based learning in their instructional practice	-0.54	-2.19	2.19
IP2G11B	Use ICT-based forms of student assessment	0.04	-1.46	1.46
IP2G11C	Use ICT for monitoring student progress	-0.01	-1.66	1.66
IP2G11G	Integrate ICT into teaching and learning	-1.97	-2.66	2.66
IP2G11H	Use subject-specific digital learning resources (e.g. tutorials, simulation)	-0.33	-2.19	2.19
IP2G11I	Use e-portfolios for assessment	1.60	-1.38	1.38
IP2G11J	Use ICT to develop authentic (real-life) assignments for students	0.94	-2.06	2.06
IP2G11K	Assess students' [computer and information literacy]	0.27	-1.59	1.59
P_EXPTCH	*Are teachers in your school expected to acquire knowledge and skills in each of the following activities?*			
IP2G11D	Collaborate with other teachers via ICT	-0.49	-2.70	2.70
IP2G11E	Communicate with parents via ICT	0.17	-1.77	1.77
IP2G11F	Communicate with students via ICT	0.32	-2.32	2.32
IP2G02I	Communicate with parents	0.80	-0.92	-0.59

School principals' reports on priorities for ICT use at schools

School principals were asked in Question 15 to indicate the priority given to different ways of facilitating the use of ICT in teaching and learning. For each item they were asked to select "high priority," "medium priority," "low priority," or "not a priority." The data from the first three items in the question were used to derive the scale *principals' reports on priorities for facilitating use of ICT - hardware* (P_PRIORH). The data from the remaining seven items were used to derive the scale of *principals' reports on priorities for facilitating use of ICT - support* (P_PRIORS). Higher values on each scale reflect higher levels of perceived priority.

Figure 12.22 depicts the results from the confirmatory factor analysis of the scaled items from the question. There was a satisfactory fit for the two-factor model, and we found a high positive correlation between the two latent factors (0.61). The reliabilities (Cronbach's alpha) of the scales for each country are presented in Table 12.41. On average, the reliability of P_PRIORH across countries was 0.79 (ranging from 0.43 to 0.90), while the average reliability of P_PRIORS across countries was 0.84 (ranging from 0.77 to 0.92). Table 12.42 records the IRT item parameters used to scale these items.

Figure 12.22: Confirmatory factor analysis of items measuring school principals' reports on priorities for ICT use at schools



Model fit indices:	Pooled sample
RMSEA	0.071
CFI	0.96
TLI	0.95

Table 12.41: Reliabilities for scales measuring school principals' reports on priorities for ICT use at schools

Country	Scale reliability (Cronbach's alpha)	
	P_PRIORH	P_PRIORS
Chile	0.89	0.90
Denmark	0.72	0.77
Finland	0.73	0.79
France	0.79	0.82
Germany	0.68	0.86
Italy	0.76	0.83
Kazakhstan	0.90	0.86
Korea, Republic of	0.88	0.92
Luxembourg	0.88	0.82
Moscow (Russian Federation)	0.83	0.82
North Rhine-Westphalia (Germany)	0.43	0.85
Portugal	0.78	0.82
United States	0.80	0.84
Uruguay	0.73	0.88
ICILS 2018 average	0.77	0.84

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.42: Item parameters for scales measuring school principals' reports on priorities for ICT use at schools

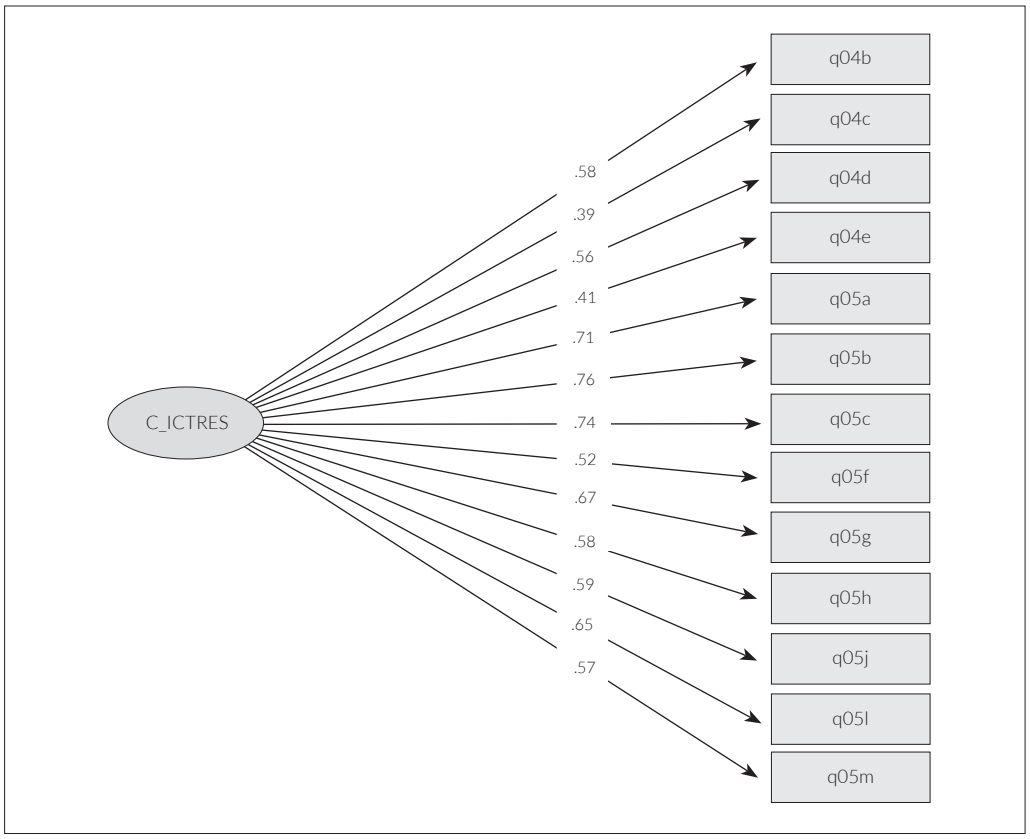
Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
P_PRIORH	*At your school, what priority is given to the following ways of facilitating the use of ICT in teaching and learning?*				
IP2G15A	Increasing the numbers of computers per student in the school	0.09	-1.40	-0.49	1.88
IP2G15B	Increasing the number of computers connected to the Internet	0.12	-0.82	-0.61	1.43
IP2G15C	Increasing the bandwidth of Internet access for the computers connected to the Internet	-0.21	-0.70	-0.69	1.39
P_PRIORS	*At your school, what priority is given to the following ways of facilitating the use of ICT in teaching and learning?*				
IP2G15D	Increasing the range of digital learning resources available for teaching and learning	-0.99	-1.95	-0.36	2.32
IP2G15E	Establishing or enhancing an online learning support platform	0.24	-1.62	-0.25	1.87
IP2G15F	Supporting participation in professional development on pedagogical use of ICT	-0.61	-1.83	-0.18	2.01
IP2G15G	Increasing the availability of qualified technical personnel to support the use of ICT	0.11	-1.22	-0.15	1.37
IP2G15H	Providing teachers with incentives to integrate ICT use in their teaching	0.26	-1.16	-0.34	1.51
IP2G15I	Providing more time for teachers to prepare lessons in which ICT is used	1.13	-1.28	-0.18	1.46
IP2G15J	Increasing the professional learning resources for teachers in the use of ICT	-0.15	-1.70	-0.38	2.08

Availability of digital resources at school

Questions 4 and 5 of the school ICT coordinator questionnaire, asked respondents to list the availability of different technology and software resources in their school with the response options: "available to teachers and students," "available only to teachers," "available only to students," or "not available." Thirteen items across the two questions were used to derive the scale *ICT coordinators reports on availability of ICT resources at school* (C_ICTRES). Higher scores correspond to greater availability of ICT-related resources.

Figure 12.23 presents the results of the confirmatory factor analysis. The one-dimensional model had a highly satisfactory fit. Table 12.43 shows the scale reliabilities (Cronbach's alpha) which had a relatively satisfactory average reliability across countries (0.74, ranging from 0.57 to 0.80). Table 12.44 shows the IRT parameters that were used to derive the scale scores.

Figure 12.23: Confirmatory factor analysis of items measuring school ICT coordinators' reports on the availability of digital resources at school



Model fit indices:	Pooled sample
RMSEA	0.048
CFI	0.89
TLI	0.87

Table 12.43: Reliabilities for scale measuring school ICT coordinators' reports on the availability of digital resources at school

Country	Scale reliability (Cronbach's alpha)
	C_ICTRES
Chile	0.80
Denmark	0.68
Finland	0.71
France	0.71
Germany	0.79
Italy	0.78
Kazakhstan	0.80
Korea, Republic of	0.79
Luxembourg	0.62
Moscow (Russian Federation)	0.78
North Rhine-Westphalia (Germany)	0.57
Portugal	0.74
United States	0.70
Uruguay	0.75
ICILS 2018 average	0.73

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.44: Item parameters for scale measuring ICT coordinators' reports on the availability of digital resources at school

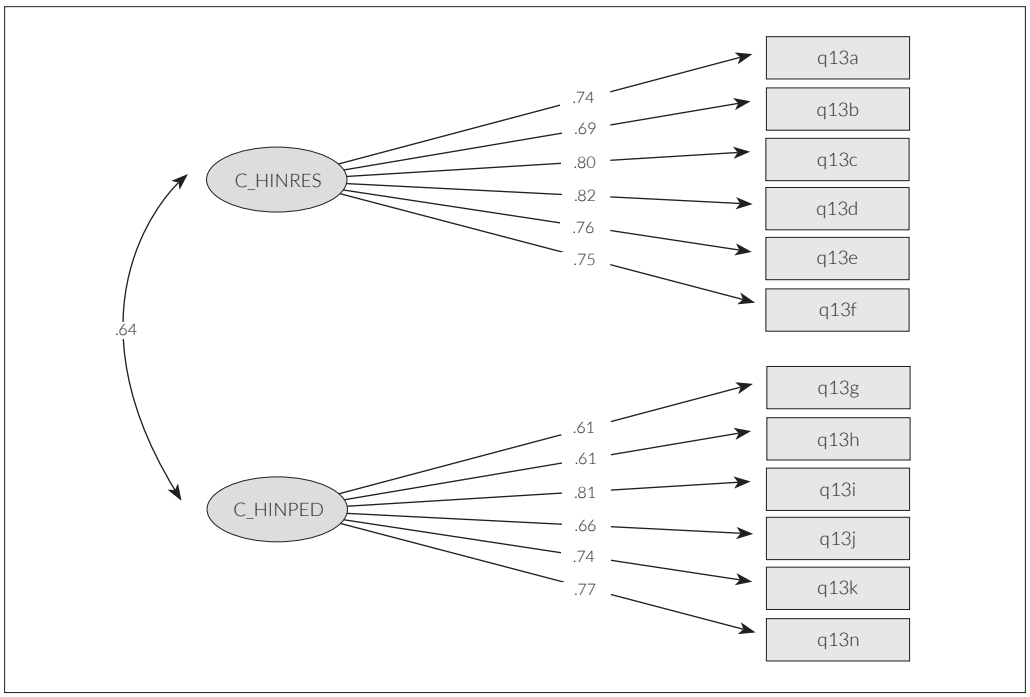
Scale or item	Question/item wording	Item parameters		
		Delta	Tau(1)	Tau(2)
C_ICTRES	*Please indicate the availability of the following technology resources in your school.*			
II2G04B	Digital learning resources that can only be used online	-0.94	1.14	-1.14
II2G04C	Access to the Internet through the school network	-1.69	0.07	-0.07
II2G04D	Access to an education site or network maintained by education authorities	-0.29	0.52	-0.52
II2G04E	Email accounts for school-related use	-0.24	-0.49	0.49
II2G05A	Practice programs or [apps] where teachers decide which questions are asked of students (e.g. [Quizlet, Kahoot], [mathfessor])	0.17	0.93	-0.93
II2G05B	Single user digital learning games (e.g. [languages online])	0.35	1.48	-1.48
II2G05C	Multi-user digital learning games with graphics and inquiry tasks (e.g. [Quest Atlantis])	1.16	1.68	-1.68
II2G05F	Video and photo software for capture and editing (e.g. [Windows Movie Maker, iMovie, Adobe Photoshop])	-0.96	1.10	-1.10
II2G05G	Concept mapping software (e.g. [Inspiration ®], [Webspiration ®])	0.57	1.39	-1.39
II2G05H	Data logging and monitoring tools (e.g. [Logger Pro]) that capture real-world data digitally for analysis (e.g. speed, temperature)	1.47	1.09	-1.09
II2G05J	A learning management system (e.g. [Edmodo], [Blackboard])	0.00	1.21	-1.21
II2G05L	e-portfolios (e.g. [VoiceThread])	0.83	1.25	-1.25
II2G05M	Digital contents linked with textbooks	-0.42	0.56	-0.56

Hindrances to the use of ICT for teaching and learning at school

In Question 13 of the ICILS 2018 ICT coordinator questionnaire, respondents were asked to indicate the extent that teaching and learning at their school is hindered by different resource-related obstacles. For each of the 14 obstacles, they could rate their impact as "a lot," "to some extent," "very little," or "not at all." Data from the first six items in the question were used to derive the scale *ICT coordinators reports on computer resource hindrances to the use of ICT in teaching and learning* (C_HINRES). Data from six of the remaining eight items were used to derive the scale *ICT coordinators reports on pedagogical resource hindrances to the use of ICT in teaching and learning* (C_HINPED). Higher scale scores corresponded to greater perceived hindrances of obstacles to the use of ICT for teaching and learning in schools.

Figure 12.24 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with the scaled items. The model had a satisfactory fit for the pooled ICILS 2018 dataset, and there was a strong correlation between latent factors (0.64).

Figure 12.24: Confirmatory factor analysis of items measuring ICT coordinators' reports on hindrances to the use of ICT for teaching and learning at school



Model fit indices:	Pooled sample
RMSEA	0.060
CFI	0.96
TLI	0.95

As evident in Table 12.45, the scale reliabilities for both scales were high. C_HINRES had an average reliability across countries of 0.81 (ranging from 0.69 to 0.88) and C_HINPED had an average reliability across countries of 0.79 (ranging from 0.68 to 0.91). Table 12.46 shows the item parameters for the scales that were used to derive the IRT scale scores.

Table 12.45: Reliabilities for scales measuring ICT coordinators' reports on hindrances to the use of ICT for teaching and learning at school

Country	Scale reliability (Cronbach's alpha)	
	C_HINRES	C_HINPED
Chile	0.85	0.83
Denmark	0.80	0.81
Finland	0.72	0.68
France	0.81	0.80
Germany	0.76	0.73
Italy	0.81	0.69
Kazakhstan	0.88	0.86
Korea, Republic of	0.86	0.91
Luxembourg	0.69	0.69
Moscow (Russian Federation)	0.79	0.88
North Rhine-Westphalia (Germany)	0.70	0.76
Portugal	0.81	0.86
United States	0.88	0.88
Uruguay	0.84	0.77
ICILS 2018 average	0.79	0.79

Notes: Benchmarking participants in *italics*. The ICILS 2018 average is based on data from the participating countries, excluding benchmarking participants.

Table 12.46: Item parameters for scales measuring ICT coordinators' reports on hindrances to the use of ICT for teaching and learning at school

Scale or item	Question/item wording	Item parameters			
		Delta	Tau(1)	Tau(2)	Tau(3)
C_HINRES	*To what extent is the use of ICT in teaching and learning at your school hindered by each of the following obstacles?*				
II2G13A	Too few computers with an Internet connection	0.76	-0.57	-0.51	1.08
II2G13B	Insufficient Internet bandwidth or speed	-0.26	-0.68	-0.18	0.85
II2G13C	Not enough computers for instruction	-0.10	-0.69	-0.40	1.08
II2G13D	Lack of sufficiently powerful computers	-0.23	-0.99	-0.08	1.07
II2G13E	Problems in maintaining ICT equipment	-0.16	-1.30	0.08	1.22
II2G13F	Not enough computer software	0.00	-1.47	0.04	1.43
C_HINPED	*To what extent is the use of ICT in teaching and learning at your school hindered by each of the following obstacles?*				
II2G13G	Insufficient ICT skills among teachers	-0.26	-1.97	-0.29	2.26
II2G13H	Insufficient time for teachers to prepare lessons	-0.24	-1.50	-0.26	1.77
II2G13I	Lack of effective professional learning resources for teachers	-0.10	-1.65	-0.18	1.83
II2G13J	Lack of an effective online learning support platform	0.35	-1.46	0.02	1.44
II2G13K	Lack of incentives for teachers to integrate ICT use in their teaching	-0.03	-1.39	-0.25	1.64
II2G13N	Insufficient pedagogical support for the use of ICT	0.28	-1.64	-0.22	1.86

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software [computer software]. Version 4. Camberwell, Australia: Australian Council for Educational Research (ACER).

Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Bollen, K. A., & Long, S. J. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*(4), 872–882.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Fabrigar, L. R., Wegener, D.T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.

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*. Cham, Switzerland: Springer. https://www.springer.com/gp/book/9783030387808

Ganzeboom, H. B. G., de Graaf, P. M., & Treiman, D. J. (1992). A standard international socioeconomic index of occupational status. *Social Science Research*, *21*, 1–56.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement equivalence in aging research. *Experimental Aging Research*, *18*, 117–144.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

International Labour Organization. (2007). *International Standard Classification of Occupations: ISCO-2008*. Geneva, Switzerland: International Labour Office.

Kaplan, D. (2009). *Structural equation modeling. Foundations and extensions*. Los Angeles, London, New Delhi, & Singapore: SAGE.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioural Research*, *32*(1), 53–76.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–49.

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). New York, Berlin, & Heidelberg: Springer.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes*. Unpublished manuscript available as MPLUS webnote. Retrieved from http://www.statmodel.com/bmuthen/articles/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (2012). *Mplus: Statistical analysis with latent variables. User's guide*. (Version 7). Los Angeles, CA: Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, vol. 2*, 113–135.

Schulz, W. (2017). Scaling of questionnaire data in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large scale education assessments* (pp. 384-410). Chichester, UK: John Wiley & Sons, Ltd.

Schulz, W., & Friedman, T. (2011). Scaling procedures for ICCS questionnaire items. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS 2009 technical report* (pp. 157–259). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., & Friedman, T. (2015). Scaling procedures for ICILS questionnaire items. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *International Computer and Literacy Information Study 2013 technical report* (pp. 177–220). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., & Friedman, T. (2018). Scaling procedures for ICCS 2016 questionnaire items. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 139–243). Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Tucker, L., & MacCallum, R. (1997). *Exploratory factor analysis*. Unpublished manuscript. Retrieved from: http://www.unc.edu/~rcm/book/factornew.htm

UNESCO. (2012). *International Standard Classification of Education: ISCED 2011*. Montreal, Canada: UNESCO-UIS.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

CHAPTER 13:

The reporting of ICILS 2018 results

Wolfram Schulz

Overview

This chapter describes the procedures used for reporting results in the ICILS 2018 international report. It illustrates the replication method used to estimate sampling variance, and how we computed the imputation variance of the computer and information literacy (CIL) and computational thinking (CT) scores. In the subsequent section, we describe how we conducted significance tests for differences between country and subgroup means or percentages, as well as between results from ICILS 2018 and 2013 for the four countries that participated in both surveys.

This chapter also explains how we conducted multiple regression analyses to explain variation in questionnaire scale scores reflecting teachers' emphasis on teaching CIL- and CT-related ICT skills, and how we estimated multilevel (hierarchical) models explaining variation in students' CIL and CT.

Estimation of sampling variance

As in the previous survey, ICILS 2018 employed two-stage cluster sampling procedures to obtain the student and teacher samples. During the first stage, schools were sampled from a sampling frame with a probability proportional to their size (see Chapter 6 for further details). During the second stage, students at the target grade were randomly sampled within schools across classrooms. Cluster sampling techniques permit an efficient and economic data collection. However, because these samples are not simple random samples and individuals within clusters tend to be more homogenous compared to the whole population, it is not appropriate to apply formulae for obtaining standard errors of sampling error for population estimates for simple random samples.

Replication (re-sampling) techniques provide tools to estimate the sampling variance of population estimates more appropriately (Gonzalez and Foy 2000; Wolter 1985). For ICILS 2018, as in the previous cycle in 2013 (see Schulz 2015), we used the jackknife repeated replication (JRR) technique to compute standard errors for population means, percentages, regression coefficients, and any other population statistic.

Generally, the JRR method for stratified samples requires pairing primary sampling units—in this survey, schools—into sampling zones (or "pseudo-strata"). Because assignment of schools to these sampling zones needs to be consistent with the sampling frame from which they were sampled, we constructed sampling zones within explicit strata. Whenever we found an odd number of schools within an explicit stratum or the sampling frame, we randomly divided the remaining school into two halves, thereby forming a sampling zone of two "quasi-schools."

Each of the countries participating in ICILS 2018 had up to 75 sampling zones (see Table 13.1), corresponding to the unpaired sample size of 150 schools. In countries where we had larger numbers of schools, we combined some schools into bigger "pseudo-schools" in order to keep the total number to 75. If a selected school was large enough to be selected with certainty, it was randomly split into two halves, which were paired. In Luxembourg, where a census of schools was surveyed, each of the 38 participating schools constituted a sampling zone and students were randomly assigned to two "quasi-schools."

Table 13.1: Number of jackknife zones in national samples

Country	Student data	Teacher data	School data
Chile	75	75	75
Denmark	72	69	72
Finland	74	73	73
France	75	65	75
Germany	75	75	75
Italy	75	74	73
Kazakhstan	75	75	75
Korea, Republic of	75	74	75
Luxembourg	38	28	35
Moscow (Russian Federation)	75	75	75
North Rhine-Westphalia (Germany)	55	54	56
Portugal	75	75	75
United States	75	75	75
Uruguay	75	62	75

Note: Benchmarking participants in *italics*.

Within each of the sampling zones, we randomly assigned one school a multiplication value of 2 and the other school a value of 0. For each of the sampling zones, we computed replicate weights. This meant that one of the paired schools had a contribution of 0, the second a double contribution, and all other schools remained the same. Replicate weights are derived by simply multiplying student, teacher, or school weights with the jackknife's multiplication value for the respective sampling zone while keeping student, teacher, or school weights for all other zones unchanged.

This process results in a replicate weight being added to the data file for each jackknife zone. Table 13.2 illustrates this procedure through a simple example featuring 24 students from six different schools (A–F) paired into three jackknife zones.

For each country sample, we computed 75 replicate weights regardless of the number of zones. In countries with fewer zones, the remaining replicate weights were set equal to the original sampling weight and therefore did not contribute to the sampling variance estimate.

Estimating the sampling variance for a statistic, μ, involves computing it once with the sampling weights for the original sample and then with each of the 75 replication weights separately. The sampling variance SV_μ estimate is computed using the formula:

$$SV_\mu = \sum_{i=1}^{75} [\mu_i - \mu_s]^2$$

Here, μ_s is the statistic μ estimated for the population through use of the original sampling weights and μ_i is the same statistic estimated by using the weights for the i^{th} of 75 jackknife replicates. The standard error SE_μ for statistic μ, which reflects the uncertainty of the estimate due to sampling, is computed as:

$$SE_\mu = \sqrt{SV_\mu}$$

The computation of sampling variance using jackknife repeated replication can be obtained for any statistic, including means, percentages, standard deviations, correlations, regression coefficients, and mean differences.

Table 13.2: Example for computation of replicate weights

ID	Student weight	School	Jackknife zone	Jackknife replicate code	Multiplication value	Replicate weight 1	Replicate weight 2	Replicate weight 3
1	5.2	A	1	0	0	0	5.2	5.2
2	5.2	A	1	0	0	0	5.2	5.2
3	5.2	A	1	0	0	0	5.2	5.2
4	5.2	A	1	0	0	0	5.2	5.2
5	9.8	B	1	1	2	19.6	9.8	9.8
6	9.8	B	1	1	2	19.6	9.8	9.8
7	9.8	B	1	1	2	19.6	9.8	9.8
8	9.8	B	1	1	2	19.6	9.8	9.8
9	6.6	C	2	1	2	6.6	13.2	6.6
10	6.6	C	2	1	2	6.6	13.2	6.6
11	6.6	C	2	1	2	6.6	13.2	6.6
12	6.6	C	2	1	2	6.6	13.2	6.6
13	7.2	D	2	0	0	7.2	0	7.2
14	7.2	D	2	0	0	7.2	0	7.2
15	7.2	D	2	0	0	7.2	0	7.2
16	7.2	D	2	0	0	7.2	0	7.2
17	4.9	E	3	1	2	4.9	4.9	9.8
18	4.9	E	3	1	2	4.9	4.9	9.8
19	4.9	E	3	1	2	4.9	4.9	9.8
20	4.9	E	3	1	2	4.9	4.9	9.8
21	8.2	F	3	0	0	8.2	8.2	0
22	8.2	F	3	0	0	8.2	8.2	0
23	8.2	F	3	0	0	8.2	8.2	0
24	8.2	F	3	0	0	8.2	8.2	0

Standard statistical software do not always include procedures for replication techniques. For ICILS 2018, we used tailored Statistical Package for the Social Sciences (SPSS) software macros (IBM Corp 2017). These results can be replicated by using the IEA International Database (IDB) Analyzer, which is generally recommended as a tool for analyzing IEA data.[1] Alternatively, analysts can use other specialized software, such as WesVar (Westat 2007), tailored applications like the SPSS Replicates Module developed by ACER,[2] or procedures built in to statistical software such as Stata (StataCorp 2013) or SAS (SAS Institute Inc. 2017).

1 The IDB Analyzer is an application that allows the user to combine and analyze data from IEA's large-scale assessments such as TIMSS, PIRLS, ICCS, and ICILS by creating SPSS or SAS Syntax, which can be used with the respecting statistic software. The application can be downloaded at https://www.iea.nl/data-tools/tools.

2 The module is an add-in component running under SPSS and offers a number of features for applying different replication methods when estimating sampling and imputation variance. The application can be downloaded at https://iccs.acer.org/iccs-2016-reports/.

Estimation of imputation variance for CIL and CT scores

The estimation of sampling variance as described above is sufficient for any analysis not involving test scores for CIL or CT. When estimating standard errors of estimates involving test scores for CIL and CT, it is important to additionally take the imputation variance into account, which provides an estimate of measurement variance (see Chapter 11 for a description of the scaling methodology for ICILS 2018 test items). Therefore, population statistics and their errors for ICILS 2018 CIL and CT scores should always be estimated using all five plausible values.

If θ is the international CIL or CT score and μ_θ^p is the statistic of interest computed on each plausible value P, then the statistic μ_θ based on all plausible values can be computed as follows:

$$\mu_\theta = \frac{1}{P}\sum_{p=1}^{P}\mu_\theta^p$$

The sampling variance SV_μ is calculated as the average of the sampling variance for each plausible value SV_μ^p:

$$SV_\mu = \frac{1}{P}\sum_{p=1}^{P}SV_\mu^p$$

Use of the P plausible values for data analysis also allows an estimation of the amount of error associated with the measurement of CIL and CT. The measurement variance or imputation variance IV_p is computed as:

$$IV_p = \frac{1}{P-1}\sum_{p=1}^{P}(\mu_\theta^p - \mu_\theta)^2$$

Here, μ_θ^p is the statistic of interest computed on each plausible value p and μ_θ is the mean statistic based on all P plausible values.

The estimate of the total variance TV_μ, consisting of sampling variance and imputation variance, can be computed as:

$$TV_\mu = SV_\mu + (1+\frac{1}{P})IV_\mu$$

The estimate of the final standard error SE_μ is equal to:

$$SE_\mu = \sqrt{TV_\mu}$$

The following formula illustrates the whole process of the computation of standard errors for a statistic (μ) based on P plausible values for 75 replicates, where μ_p^i is the statistic for the i^{th} replicate of the pth plausible value, and μ_p is the statistic for the p^{th} plausible value with the original sampling weights:

$$SE_\mu = \sqrt{\left[\sum_{p=1}^{P}\left(\sum_{i=1}^{75}(\mu_p^i - \mu_p)^2\right)/P\right] + \left[(1+\frac{1}{P})\frac{\sum_{p=1}^{P}(\mu_p - \bar{\mu}_p)^2}{P-1}\right]}$$

Table 13.3 shows the average scale scores for CIL as well as their sampling and overall standard errors, while Table 13.4 displays the corresponding information for CT. The tables also record the number of students that were assessed in each country. The comparison between sampling and combined standard error shows that for both assessment domains most of the error was due to sampling and that (at the level of national samples) only a relatively small proportion was attributable to measurement error.

Table 13.3: National averages for CIL with standard deviations, sampling, and overall errors

Country	Average CIL score	Sampling error	Combined standard error	Number of assessed students
Chile	476	3.71	3.73	3092
Denmark	553	2.00	2.04	2404
Finland	531	2.96	2.98	2546
France	499	2.26	2.35	2940
Germany	518	2.81	2.95	3655
Italy	461	2.69	2.78	2810
Kazakhstan	395	5.33	5.37	3371
Korea, Republic of	542	2.95	3.05	2875
Luxembourg	482	0.68	0.83	5401
Moscow (Russian Federation)	549	2.21	2.25	2852
North Rhine-Westphalia (Germany)	515	2.55	2.63	1991
Portugal	516	2.58	2.59	3221
United States*	519	1.86	1.89	6790
Uruguay	450	4.19	4.29	2613

Notes: Benchmarking participants in *italics*.
* Countries not meeting sample participation requirements.

Table 13.4: National averages for CT with standard deviations, sampling, and overall errors

Country	Average CT score	Sampling error	Combined standard error	Number of assessed students
Denmark	527	2.29	2.32	2404
Finland	508	3.32	3.37	2546
France	501	2.31	2.38	2940
Germany	486	3.54	3.63	3655
Korea, Republic of	536	4.26	4.42	2875
Luxembourg	460	0.86	0.89	5401
North Rhine-Westphalia (Germany)	485	2.87	2.96	1991
Portugal	482	2.48	2.51	3221
United States*	498	2.48	2.54	6790

Notes: Benchmarking participants in *italics*.
* Countries not meeting sample participation requirements.

Reporting of differences

Differences in population estimates between and within countries

We considered differences between two score averages (or percentages) *a* and *b* significant ($p < 0.05$) when the test statistic *t* was greater than the critical value, 1.96. We calculated *t* by dividing the difference by its standard error, SE_{dif_ab}:

$$t = \frac{(a-b)}{SE_{dif\_ab}}$$

In the case of differences between score averages from independent samples (evident, for example, with respect to comparisons of two country averages), the standard error of the difference SE_{dif_ab} can be computed as:

$$SE_{dif\_ab} = \sqrt{SE_a^2 + SE_b^2}$$

Here, SE_a and SE_b are the standard errors of the means from the two independent samples *a* and *b*.

The formula for calculating the standard error provided above is only suitable when the subsamples being compared are independent. Because subgroups (e.g., gender groups) within countries are typically not independent samples, we derived the difference between statistics for subgroups of interest and the standard error of the difference by using jackknife repeated replication that involved the following formula:

$$SE_{dif\_ab} = \sqrt{\left[\sum_{i=1}^{75}((a^i - b^i) - (a - b))^2\right]}$$

Here, *a* and *b* represent the averages (or percentages) in each of the two subgroups for the fully weighted sample, and a^i and b^i are those for the replicate samples.

In the case of differences in CIL and CT scores between dependent subsamples, we calculated the standard error of the differences with ($P = 5$) plausible values by using this formula:

$$SE_{dif\_ab} = \sqrt{\left[\sum_{p=1}^{P}\left(\sum_{i=1}^{75}((a_p^i - b_p^i) - (a_p - b_p))^2\right)/P\right] + \left[(1 + \frac{1}{P})\frac{\sum_{p=1}^{p}((a_p - b_p) - (\bar{a}_p - \bar{b}_p))^2}{P-1}\right]}$$

Here, a_p and b_p represent the weighted subgroup averages in groups *a* and *b* for each of the P plausible values, a_p^i and b_p^i are the subgroup averages within replicate samples for each of the P plausible values, and $\bar{a}_p$ and $\bar{b}_p$ are the means of the two weighted subgroup averages across the P plausible values.

Comparisons between countries and ICILS averages

The standard error of the ICILS 2018 average SE_{μ_2018} was calculated based on the respective standard error for each of the national statistics (SE_k) and the number (N) of countries meeting IEA sample participation requirements that were included in the average (11 for student survey results, 7 for teacher survey results):

$$SE_{\mu\_2018} = \frac{\sum_{k=1}^{N} SE_k^2}{N}$$

When comparing the country means *c* with the overall ICILS 2018 average *i*, we had to account for the fact that the country being considered had contributed to the international standard error. We did this by calculating the standard error SE_{dif_ic} of the difference between the overall ICILS 2018 average and an individual country average as:

$$SE_{dif\_ic} = \frac{\sqrt{((N-1)^2 - 1)SE_c^2 + \sum_{k=1}^{N} SE_k^2}}{N}$$

Here, SE_c is the sampling standard error for country *c* and SE_k is the sampling error for k^{th} of *N* participating and reported countries. We used this formula for determining the statistical significance of differences due to sampling error between countries and the ICILS 2018 averages of all questionnaire percentages or scale point averages throughout the ICILS 2018 reports.

When comparing the CIL and CT score averages of a country with the overall ICILS 2018 average, it was necessary to also account for the imputation component of standard errors for countries into account. The imputation variance component of standard errors $SE^2_{i\_dif\_icp}$ was given as:

$$SE^2_{i\_dif\_icp} = \sqrt{(1+\frac{1}{P})\, var(d_1,\ldots,d_p,\ldots,d_5)}$$

Here, d_p is the difference between the overall ICILS 2018 average and the country mean for the plausible value p.

The sampling error for CIL and CT scores was calculated as follows:

$$SE_{dif\_icp} = \frac{\sqrt{((N-1)^2-1)\left(\frac{1}{5}\sum_{p=1}^{5} SE^2_{cp} + \sum SE^2_k\right) + \sum_{k=1}^{N}\left(\frac{1}{5}\sum_{p=1}^{5} SE^2_{kp}\right)}}{N}$$

Here, SE_{cp} is the sampling standard error for country c and plausible value p, and SE_{kp} is the sampling error for plausible value p in the k^{th} of N participating and reported countries.

We computed the final standard error (SE_{dif_icp}) of the difference between national CIL and CT country test scores and the ICILS 2018 averages as:

$$SE_{a\_dif\_icp} = \sqrt{SE^2_{dif\_icp} + SE^2_{i\_dif\_icp}}$$

Comparisons between benchmarking participants and ICILS averages

When comparing averages for benchmarking participants, Moscow (Russian Federation) constituted an independent student sample in relation to the ICILS 2018 average while North Rhine-Westphalia was a sub-sample of the German national student sample (representing 22.5% of the corresponding student population). Therefore, two different formulas had to be applied. For the teacher survey, the German teacher survey did not meet IEA sample participation requirements and was therefore not included in the ICILS 2018 average for results from the teacher survey. Therefore, North Rhine-Westphalia, which had met sample participation requirements as a benchmarking participant, constituted an independent sample in relation to the ICILS 2018 average for teacher results.

Standard errors for differences between ICILS 2018 averages and Moscow (Russian Federation), both for student and teacher results, as well as North Rhine-Westphalia (Germany) with regard to teacher results were computed with the following formula:

$$SE_{dif\_ic} = \sqrt{SE^2_{bp} + SE^2_{\mu\_2018}}$$

Here, SE_{bp} is the standard error for the result from the benchmarking participant, and SE_{μ_2018} is the standard error for the ICILS 2018 average.

For differences between student (or school) survey results in North Rhine-Westphalia (Germany) and the ICILS 2018 average, the standard error SE_{dif_StNRW} was computed as follows:

$$SE_{dif\_StNRW} = \frac{\sqrt{((N-1)^2-0.225)SE^2_{StNRW} + \sum_{p=1}^{5} SE^2_{cp} + \sum_{k=1}^{N} + SE^2_k}}{N}$$

Here, SE_{StNRW} represents the standard error for student survey estimate from North Rhine-Westphalia and SE_k is the sampling error for k^{th} of N participating countries. The correction for the contribution of the data from this benchmarking participant was set to 0.225 (instead of 1) as its student population was equivalent to 22.5 percent of the overall German population.

Comparisons between CIL results in 2018 and 2013

For those countries that had participated in the previous cycle, the ICILS 2018 international report also included comparisons of test results for CIL between ICILS 2013 and 2018. Because the process of equating the CIL scores across the cycles introduced some additional error into the calculation of any test statistic, we added an equating error term to the formula for the standard error of the difference between country averages.

When testing the difference of a statistic between the two assessments, we computed the standard error of the difference as follows:

$$SE(\mu_{(ICILS18)} - \mu_{(ICILS13)}) = \sqrt{SE^2_{\mu ICILS18} + SE^2_{SE^2 \mu ICILS13} + EqErr^2}$$

Here, μ can be any statistic in units on the equated ICILS scale (mean, percentile, gender difference, but not percentages) and $SE_{\mu ICILS18}$ and $SE_{\mu ICILS13}$ are the respective standard errors of this statistic from the two surveys. *EqErr* denotes the equating error that reflects the uncertainty in the link between both assessments, which was equal to 3.9 score points for the CIL scale (see Chapter 11 for the calculation of the respective equating error).

To report the significance of differences between ICILS 2018 and 2013 for percentages of students with CIL scores at or above Level 2 (see Chapter 1 for a description of these levels), it was not possible to use the estimated equating error in CIL score points. Therefore, we applied the following replication method to estimate the corresponding equating errors.

To estimate the standard error of the percentage at or above the cut-point that defines the threshold between Levels 1 and 2 (492), within each participating country a number of *n* replicate cut-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 (3.9). Percentages of students at or above each replicate cut-point (ρ_n) were computed for each of these replicated thresholds and the equating error for each participating country was estimated as:

$$EquErr_{\rho\_country} = \sqrt{\frac{(\rho_n - \rho_o)}{N_{rep}}}$$

Here, ρ_o is the observed percentage of students at or above Level 2. We used 1000 replicate samples (N_{rep}) for these computations. The equating errors for the national percentages of students at or above Level 2 were estimated as 1.95 for Chile, 1.23 for Denmark, 1.45 for Germany, and 1.18 for Korea.

Within each participating country, the standard errors for the differences between percentages at or above proficient levels were calculated based on the standard error on the percentage at or above Level 2 in 2018 ($SE_{\rho ICILS18}$), the standard error for the corresponding estimate in 2013 ($SE_{\rho ICILS13}$), and the estimated equating error ($EquError_{\rho country}$) as follows:

$$SE(\rho_{(ICILS18)} - \rho_{(ICILS13)}) = \sqrt{SE^2_{\rho ICILS18} + SE^2_{\rho ICILS13} + EquErr^2_{\rho country}}$$

Multiple regression modeling of teacher data

When reporting ICILS 2018 data, we also used single-level multiple regression models to explain variation in the questionnaire scale scores reflecting teachers' emphasis on teaching CIL and CT, respectively. Predictor variables were teachers' ICT self-efficacy, positive perceptions of pedagogical ICT use, perceptions of higher levels of teacher collaboration, and reports on higher levels of availability of ICT resources at their school and teachers' experience with using ICT during lessons.

Multiple regression models (see, for example, Pedhazur 1997) were estimated as:

$$Y_i = \beta_0 + \beta_i X_i^n + \varepsilon_i$$

Here, for sample of *i* teachers we regressed our criterion variables Y_i (teachers' emphasis on teaching CIL or CT) on a vector of predictors X_i^n with its corresponding vector of regression coefficients β_i, where β_0 denotes the intercept, and ε_i represents the unexplained part of the model (residual).

We reported the unstandardized regression coefficients and the variance explained by the model to show the effects for each predictor and the overall explanatory power of the model. To estimate standard errors for the multiple regression model parameters, we employed jackknife repeated replication using tailored SPSS macros, which can be exactly replicated with the IEA IDB Analyzer.

Table 13.5 shows the numbers of all assessed teachers in each country, of teachers included in each of the multiple regression analyses, as well as the weighted percentages of teachers with valid data for all variables in each of the two estimated models.

Table 13.5: ICILS 2018 teachers included in multiple regression analyses of teachers' emphasis on teaching CIL- and CT-related skills

Country	Multiple regression analysis of teachers' emphasis on CIL-related skills		Multiple regression analysis of teachers' emphasis on CT-related skills	
	Total number of teachers in analysis	Weighted percentage of teachers in analysis	Total number of teachers in analysis	Weighted percentage of students in analysis
Chile	1625	96	1622	96
Denmark	1081	97	1078	97
Finland	1797	97	1795	97
France	1405	96	1400	96
Germany	2212	94	2207	94
Italy	2534	97	1721	97
Kazakhstan	2534	96	2535	96
Korea, Republic of	2068	97	2048	96
Luxembourg	473	96	472	96
Moscow (Russian Federation)	2189	97	2185	97
North Rhine-Westphalia (Germany)	1398	94	1400	95
Portugal	2743	97	2750	98
United States*	3103	96	3103	96
Uruguay	1178	90	1176	90

Notes: Benchmarking participants in *italics*.
* Countries not meeting sample participation requirements.

Across the participating countries, we observed an average percentage of teachers in the sample with valid data for all variables of 96 percent. National average percentages of teachers with valid data for all variables ranged from 90 percent in Uruguay to 98 percent in Portugal.[3]

3 Readers should note that when applying models with a larger number of predictor variables, it is likely that the proportion of missing values increases when applying a (list-wise) exclusion of respondents with omitted data for any of the variables in the analysis.

Hierarchical linear modeling to explain variation in students' CIL and CT

To review which factors are associated with variation in CIL and CT within and across schools within participating countries, we estimated within each country hierarchical (or multilevel) linear regression models (Raudenbush and Bryk 2002) in which students were nested within schools. Predictor variables included variables reflecting students' personal and social background, ICT-related variables at the student level, and ICT-related factors at the school level.

A hierarchical regression model with i students nested in j clusters (schools) can be estimated as:

$$Y_{ij} = \beta_0 + \beta_{ij} X_{ij}^n + \beta_j X_j^m + U_{0j} + \varepsilon_{ij}$$

Here, Y_{ij} is the criterion variable, β_0 is the intercept, X_{ij}^n is a vector of student-level variables, with its corresponding vector of regression coefficients β_{ij}, and Xmj is a school-level variable with its corresponding vector of regression coefficients β_j. U_{0j} is the residual term at the level of the cluster (school), and ε_{ij} is the student-level residual. Both residual terms are assumed to have a mean of 0 and variance that is normally distributed at each level.

The explained variance in hierarchical linear models has to be estimated for each level separately, with the estimate based on a comparison of each prediction model with the baseline ("null") model (or ANOVA model) without any predictor variables.

We estimated the null model, from which we excluded students with missing data after completing "missing treatment" (see section on missing treatment below) as:

$$Y_{ij} = \beta_0 + U_{0j(null)} + \varepsilon_{ij\,(null)}$$

The residual term $U_{0j(null)}$ provides an estimate of the variance in Y_{ij} between *j* clusters, and $\varepsilon_{ij\,(null)}$ is an estimate of the variance between *i* students within clusters. The intra-class correlation *IC*, which reflects the proportion of variance between clusters (in our case, schools), can be computed from these estimates as:

$$IC = \frac{U_{0j\,(null)}}{U_{0j(null)} + \varepsilon_{ij\,(null)}}$$

Based on the estimates of variances at school and student level derived from the null model, we computed the explained variance at the school level EV_j as:

$$EV_j = \left(1 - \frac{U_{0j}}{U_{0j\,(null)}}\right) \times 100$$

We computed the explained variance at the student level EV_{ij} as:

$$EV_{ij} = \left(1 - \frac{\varepsilon_{ij}}{\varepsilon_{ij\,(null)}}\right) \times 100$$

Because multilevel modeling takes the hierarchical structure of the cluster sample into account and we used plausible values as estimates of students' CIL and CT in our models, the reported multilevel standard errors reflect both sampling and imputation errors.

National data were weighted with normalized school-level and [within-school] student-level weights, where the sum of within- and between-school weights is equal to the sample size (Asparouhov 2006). Normalized (or scaled) within-school student-level weights ω_{ij}^* were calculated as:

$$\omega_{ij}^* = \omega_{ij} \frac{n_j}{\sum_i \omega_{ij}}$$

Here, n_j represents the cluster size (equal to the number of students with valid data within each school) and ω_{ij} the original within-school student-level weights. Normalized (or scaled) school-level weights ω_j^* were computed as:

$$\omega_j^* = \omega_j \frac{n}{\sum_{i,j} \omega_i^* \omega_j}$$

Here, n is the total sample size and ω_j denotes the original school level weights.

We used the software package MPlus (Version 7, see Muthén and Muthén 2012) to estimate all hierarchical models. Even though Luxembourg had met the sample participation requirements for the student survey and 38 out of 41 of its schools with target grade students participated, we excluded their data from the analyses; the low number of schools would have resulted in a greatly reduced statistical power and a rather limited precision of estimation of school level effects. Results from the United States, which did not meet IEA sample participation requirement, were reported separately and should be interpreted with caution.

As is customary when applying multivariate analyses, we observed increases in the proportions of missing data when including more variables in the model. To account for higher proportions of missing responses for some of the variables with higher percentages of missing values, the analysis included a "dummy variable adjustment" for these data (see Cohen and Cohen 1975). For each of these variables, we assigned mean or median values to cases with missing data and added dummy indicator variables (with 1 indicating a missing value and 0 non-missing values) to the analysis.

At the student level we observed higher proportions of missing values for the scale reflecting use of general ICT applications in class and the scales measuring students' perceptions of learning of CIL- or CT-related skills (which were included as predictor variables for explaining CIL and CT respectively). Given that information from teachers tended to be either not missing or missing (in almost all cases) for both predictor variables (average years of teachers' experience with ICT use during lessons and teachers' reports on students' ICT use for class activities) from this survey, only one missing indicator was created to indicate missing teacher data for both variables. Two further missing indicators were used for missing school principal data regarding schools' expectations of teacher communication via ICT, and for missing ICT coordinator data about the availability of ICT resources at school.

Table 13.6 shows the coefficients for missing indicators included in the multilevel analyses of CIL and CT in each country. Student-level missing indicators tended to be negatively related with CIL and CT respectively while patterns were less consistent for school-level predictors derived from teacher, principal, and ICT coordinator surveys. In countries where there were no schools with missing information from either school or teacher questionnaires, it was not necessary to apply any treatment. Consequently, no missing indicator coefficients are displayed in this table (as in Korea for school principal and ICT coordinator questionnaire data, in Kazakhstan for teacher information, and in Moscow for all school-level data).

Table 13.7 shows the numbers of students included in the multilevel analyses, as well as the respective weighted percentages of students with valid data for all variables in the model.

For the multilevel analyses of CIL, 92 percent of students, on average across ICILS 2018 countries meeting sample participation requirements, had valid data for all variables included in the model. In Germany and Uruguay, however, only 84 and 83 percent (respectively) of the weighted sample had valid data for inclusion in the analyses of CIL, and consequently their results should be interpreted with due caution. For the analysis of variation in CT, on average 94 percent of students had valid data for all variables in the model. However, in Germany, only 84 percent of the weighted sample could be included in these analyses with similar caveats that should be observed when interpreting the respective results.

Table 13.6: Coefficients of missing indicators in multilevel analysis of CIL and CT data

Country	Missing indicator for use of general ICT applications in class		Missing indicator for CIL- and CT-learning respectively		Missing indicator for school's expectations of teacher communication via ICT		Missing indicator for ICT resources at school		Missing teacher data indicator	
	CIL	CT	CIL	CT	CIL	CT	CIL	CT	CIL	CT
Chile	**-32.6** (15.1)		-24.8 (17.0)		-8.8 (15.8)		-9.1 (13.3)		-8.2 (32.1)	
Denmark	**-33.1** (16.0)	**-42.5** (18.5)	-15.4 (22.2)	-17.9 (12.3)	-3.4 (13.0)	2.0 (8.5)	2.2 (8.0)	-1.2 (5.3)	-2.3 (9.7)	12.4 (12.8)
Finland	**-87.2** (25.5)		**-69.7** (31.7)		-6.0 (7.2)		19.2 (15.8)		-7.0 (6.4)	
France	**-39.1** (9.2)	**-44.4** (10.3)	**-35.6** (10.0)	**-28.2** (8.1)	-2.4 (10.0)	-1.3 (9.3)	-0.1 (6.7)	0.1 (8.9)	**-16.6** (6.9)	-7.8 (7.3)
Germany	-28.1 (18.1)	-38.6 (21.3)	-8.3 (9.6)	5.7 (7.1)	13.0 (19.0)	4.1 (13.7)	2.9 (17.2)	-13.9 (14.9)	-21.3 (21.4)	-16.6 (9.5)
Italy	**-50.0** (20.6)		-49.9 (17.5)		-0.7 (9.7)		-4.8 (15.7)		**-101.0** (16.1)	
Kazakhstan	-33.0 (22.4)		-55.5 (29.9)		6.4 (21.2)		-21.3 (24.1)		N/A	
Korea, Republic of	-48.7 (61.0)	**-147.5** (58.8)	34.6 (84.7)	-7.5 (51.0)	N/A	N/A	N/A	N/A	14.3 (14.2)	10.5 (20.1)
Luxembourg	**-41.0** (6.2)	**-60.1** (8.3)	**-32.6** (6.5)	**-24.7** (6.5)	3.9 (10.9)	-2.9 (7.1)	1.7 (12.5)	**12.8** (6.3)	0.8 (13.6)	-1.1 (6.4)
Moscow (Russian Federation)	**-53.1** (20.4)		-16.6 (12.8)		N/A	N/A	N/A	N/A	N/A	N/A
North Rhine-Westphalia (Germany)	-8.1 (12.6)	-25.5 (13.2)	**-36.2** (11.7)	8.4 (7.9)	**-56.9** (21.7)	-37.0 (21.1)	6.3 (10.5)	15.7 (12.3)	17.2 (18.7)	-1.6 (16.7)
Portugal	**-61.1** (17.2)	**-57.3** (22.5)	32.0 (16.6)	-15.9 (15.4)	-26.5 (21.4)	-7.8 (13.9)	-0.3 (12.0)	**-27.8** (10.5)	-14.5 (10.3)	8.8 (10.1)
United States*	**-26.3** (10.0)	**-45.1** (10.8)	**-29.8** (10.7)	-12.6 (6.9)	-6.3 (10.8)	-12.5 (9.3)	6.8 (8.0)	13.3 (12.2)	**-29.7** (11.7)	**-53.2** (16.1)
Uruguay	**-27.1** (8.7)		**-33.5** (7.8)		-6.9 (11.5)		3.2 (13.5)			

Notes: Benchmarking participants in *italics*.
* Countries not meeting sample participation requirements. N/A = Full information available

Table 13.7: ICILS 2018 students included in multilevel analyses of variation in CIL and CT

Country	CIL		CT	
	Number of students in analysis	Weighted % of students in analysis	Number of students in analysis	Weighted % of students in analysis
Chile	2888	93		
Denmark	2319	97	2319	97
Finland	2453	97	2453	97
France	2686	92	2686	92
Germany	2983	84	2983	84
Italy	2617	93		
Kazakhstan	3169	95		
Korea, Republic of	2786	97	2786	97
Moscow (Russian Federation)	2718	95		
North Rhine-Westphalia (Germany)	1562	79	1562	79
Portugal	3081	96	3081	96
United States*	5829	86	5829	86
Uruguay	2126	83		

Notes: Benchmarking participants in *italics*.
* Countries not meeting sample participation requirements.

References

Asparouhov, T. (2006). General multilevel modeling with sampling weights. *Communications in Statistics: Theory and Methods, 35*(3), 439-460.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gonzalez, E.J., & Foy, P. (2000). Estimation of sampling variance. In M. O. Martin, K. D. Gregory, & S.E. Semler (Eds.), *TIMSS 1999: Technical report*. Chestnut Hill, MA: Boston College.

IBM Corp. (2017). IBM SPSS Statistics for Windows, Version 25.0 [statistical software]. Armonk, NY: Author.

Muthén, L.K., & Muthén, B.O. (2012). *Mplus: Statistical analysis with latent variables. User's guide*. (Version 7). Los Angeles, CA: Muthén & Muthén.

Pedhazur, E.J. (1997). *Multiple regression in behavioral research: explanation and prediction* (3rd ed.). Ford Worth, TX: Harcourt Brace College.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publishers.

Schulz, W. (2015). Reporting of ICILS results. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt, (Eds.) *International Computer and Literacy Information Study 2013 technical report* (pp. 221-232). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA). SAS Institute Inc. (2017). SAS/STAT®14.3 user's guide [statistical software]. Cary, NC: Author.

StataCorp. (2013). Stata user's guide. Release 13 [statistical software]. College Station, TX: Author.

Westat. (2007). WesVar ®4.3: User's guide [computer software]. Rockville, MD: Author.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York, NY: Springer.

APPENDIX A:

Organizations and individuals involved in ICILS 2018

International study center

The international study center is located at the Australian Council for Educational Research (ACER). ACER is responsible for designing and implementing the study in close cooperation with IEA.

Staff at ACER

Julian Fraillon, *research director*
John Ainley, *project coordinator*
Wolfram Schulz, *assessment coordinator*
Tim Friedman, *project researcher*
Daniel Duckworth, *test developer*
Melissa Hughes, *test developer*
Laila Helou, *quality assurer*
Alex Daraganov, *data analyst*
Renee Kwong, *data analyst*
Leigh Patterson, *data analyst*
Louise Ockwell, *data analyst*
Katja Bischof, *project researcher*

International Association for the Evaluation of Educational Achievement (IEA)

IEA provided overall support in coordinating and implementing ICILS 2018. IEA Amsterdam, the Netherlands, was responsible for membership, translation verification, quality control, and the publication and wider dissemination of the report. IEA Hamburg, Germany, was mainly responsible for managing field operations, sampling procedures, and data processing.

Staff at IEA Amsterdam

Dirk Hastedt, *executive director*
Andrea Netten, *director IEA Amsterdam*
Roel Burgers, *financial director*
Michelle Djekić, *research and liaison officer (former)*
Sandra Dohr, *junior research officer*
David Ebbs, *senior research officer*
Sive Finlay, *head of communications (former)*
Isabelle Gémin, *senior financial officer*
Mirjam Govaerts, *public relations and events officer*
Gina Lamprell, *junior publications officer*
Jennifer Ross, *media and outreach officer*
Jasmin Schiffer, *graphic designer*
Jan-Philipp Wagner, *junior research officer*
Gillian Wilson, *senior publications officer*

Staff at IEA Hamburg

Juliane Hencke, *director IEA Hamburg*
Heiko Sibberns, *director IEA Hamburg (former)*
Ralph Carstens, *senior research advisor*
Sebastian Meyer, *ICILS international data manager*
Michael Jung, *ICILS international data manager (former)*
Ekaterina Mikheeva, *ICILS deputy international data manager*
Lars Borchert, *ICILS deputy international data manager (former)*
Sabine Meinck, *head of research and analysis, and sampling units*
Sabine Tieck, *research analyst (sampling)*
Sabine Weber, *research analyst (sampling)*
Karsten Penon, *research analyst (sampling)*
Duygu Savaşcı, *research analyst (sampling)*
Oriana Mora, *research analyst*
Adeoye Oyekan, *research analyst*
Hannah Köhler, *research analyst*
Lorelia Lerps, *research analyst*
Rea Car, *research analyst*
Clara Beyer, *research analyst*
Yasin Afana, *research analyst*
Guido Martin, *head of coding unit*
Katharina Sedelmayr, *research analyst (coding)*
Deepti Kalamadi, *programmer*
Maike Junod, *programmer*
Limiao Duan, *programmer*
Devi Prasath, *programmer (former)*
Bettina Wietzorek, *meeting and seminar coordinator*

RM Results

RM Results was responsible for developing the software systems underpinning the computer-based student assessment instruments for the main survey. This work included development of the test and questionnaire items, the assessment delivery system, and the web-based translation, scoring, and data-management modules.

RM Results

Mike Janic, *managing director*
Stephen Birchall, *deputy CEO*
Erhan Halil, *product development manager*
Rakshit Shingala, *team leader*
James Liu, *analyst programmer*
Nilupuli Lunuwila, *analyst programmer*
Richard Feng, *analyst programmer*
Stephen Ainley, *quality assurance*
Ranil Weerasinghe, *quality assurance*
Grigory Loskutov, *IT coordinator*

ICILS sampling referee

Marc Joncas was the sampling referee for the study. He provided invaluable advice on all sampling-related aspects of the study.

National research coordinators

The national research coordinators 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.

Chile
Carolina Leyton
María Victoria Martinez
Tabita Nilo
National Agency for Educational Quality

Denmark
Jeppe Bundsgaard
Danish School of Education, Aarhus University

Finland
Kaisa Leino
Finnish Institute for Educational Research, University of Jyväskylä

France
Marion Le Cam
Ministry of National Education

Germany and North Rhine-Westphalia (Germany)
Birgit Eickelmann
Institute for Educational Science, University of Paderborn

Italy
Elisa Caponera
Riccardo Pietracci
INVALSI (Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione)
Gemma De Sanctis (until May 2018)
MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca)

Kazakhstan
Aigerim Zuyeva
Ruslan Abrayev
Department for International Comparative Studies, Ministry of Education and Science

Luxembourg
Catalina Lomos
Luxembourg Institute of Socio-Economic Research (LISER)

Moscow (Russian Federation)
Elena Zozulia
Moscow Center for Quality of Education

Portugal
Vanda Lourenco
IAVE, IP—Institute of Educational Evaluation

Republic of Korea
Sangwook Park
Kyongah Sang
Korea Institute for Curriculum and Evaluation

United States
Lydia Malley
Linda Hamilton
National Center for Education Statistics, US Department of Education

Uruguay
Cristobal Cobo
Center for Research—Ceibal Foundation

Cecilia Hughes
Evaluation and Monitoring Department at Plan Ceibal

APPENDIX B:

Characteristics of national samples

For each educational system participating in ICILS 2018, this appendix describes population coverage, exclusion categories, stratification variables, and any deviations from the general ICILS 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 can differ. It is particularly common that a school counts as participating in the student survey, but not in the teacher survey; however, the reverse scenario is also possible. If the school participation status in both parts of ICILS 2018 differs, the figures are displayed in two separate tables. If the status counts are identical in both parts, the results are displayed in one combined table.

Numbers in brackets refer to the number of categories of specific stratification variables.

B.1 Chile

- School level exclusions consisted of schools for children with special educational needs, very small schools (less than six students in the target grade) and geographically inaccessible schools. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by school type (grade 8 and 9, grade 8 only), school administration type (public, private-subsidized, private), and urbanization (rural, urban), resulting in 10 explicit strata.
- Implicit stratification was applied by national assessment performance group for mathematics (four levels), giving a total of 36 implicit strata.
- The sample was disproportionally allocated to explicit strata.
- Schools were oversampled to allow for better estimates for private and rural schools.
- Small schools were selected with equal probabilities.

Table B.1.1: Allocation of student sample in Chile

School participation status—student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Grade 8 & 9 – Public – Rural	4	0	4	0	0	0
Grade 8 & 9 – Public – Urban	10	0	9	1	0	0
Grade 8 & 9 – Private subsidized – Rural	4	0	4	0	0	0
Grade 8 & 9 – Private subsidized – Urban	20	0	17	2	1	0
Grade 8 & 9 – Private – Urban & rural	46	0	38	3	5	0
Grade 8 only – Public – Rural	30	0	29	1	0	0
Grade 8 only – Public – Urban	28	0	27	1	0	0
Grade 8 only – Private subsidized – Rural	12	0	12	0	0	0
Grade 8 only – Private subsidized – Urban	22	1	21	0	0	0
Grade 8 only – Private – Urban	4	0	2	1	0	1
Total	**180**	**1**	**163**	**9**	**6**	**1**

Note: No schools with student participation rate below 50% were found.

Table B.1.2: Allocation of teacher sample in Chile

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Grade 8 & 9 – Public – Rural	4	0	4	0	0	0
Grade 8 & 9 – Public – Urban	10	0	9	0	0	1
Grade 8 & 9 – Private subsidized – Rural	4	0	4	0	0	0
Grade 8 & 9 – Private subsidized – Urban	20	0	17	1	1	1
Grade 8 & 9 – Private – Urban & rural	46	0	37	3	5	1
Grade 8 only – Public – Rural	30	0	28	1	0	1
Grade 8 only – Public – Urban	28	0	27	1	0	0
Grade 8 only – Private subsidized – Rural	12	0	12	0	0	0
Grade 8 only – Private subsidized – Urban	22	1	21	0	0	0
Grade 8 only – Private – Urban	4	0	2	1	0	1
Total	**180**	**1**	**161**	**7**	**6**	**5**

Note: Four schools with teacher participation rate below 50% were found.

B.2 Denmark

- School level exclusions consisted of schools for children with special education needs, treatment centers, schools with less than five students in the target grade, and German, English, and Waldorf schools. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- No explicit stratification was performed.
- Implicit stratification was applied by assessment score, giving a total of five implicit strata.
- Small schools were selected with equal probabilities.

Table B.2.1: Allocation of student sample in Denmark

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Denmark	150	0	114	25	4	7
Total	**150**	**0**	**114**	**25**	**4**	**7**

Note: No schools with student participation rate below 50% were found.

Table B.2.2: Allocation of teacher sample in Denmark

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Denmark	150	0	109	25	4	12
Total	**150**	**0**	**109**	**25**	**4**	**12**

Note: Two schools were regarded as non-participating because the within-school participation rate was below 50%.

B.3 Finland

- School level exclusions consisted of schools for children with special education needs and schools with instruction language not Finnish or Swedish. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by region (5), urbanization (urban, semi-urban, rural), and language (2), resulting in nine explicit strata.
- Implicit stratification was applied by region (4), and urbanization (urban, semi-urban, rural), giving a total of 17 implicit strata.
- School sample overlap between ICILS 2018 and TALIS 2018[1]: Both samples were drawn at once using minimum overlap control. The samples were proportionally allocated to explicit strata. All schools have been selected with equal probabilities.

Table B.3.1: Allocation of student sample in Finland

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools	Excluded schools
			Sampled	First replacement	Second replacement		
Helsinki/Uusimaa – Urban & semi-urban & rural	36	0	35	0	0	1	0
Southern – Urban & semi-urban	26	0	26	0	0	0	0
Southern – Rural	6	0	6	0	0	0	0
Western – Urban & semi-urban	28	1	25	0	0	1	1
Western – Rural	6	0	5	0	0	0	1
Northern & Eastern – Urban & semi-urban	28	1	26	1	0	0	0
Northern & Eastern – Rural	10	0	10	0	0	0	0
Swedish speaking – No Aland	8	0	8	0	0	0	0
Swedish speaking – Aland	2	0	2	0	0	0	0
Total	**150**	**2**	**143**	**1**	**0**	**2**	**2**

Note: No schools with student participation rate below 50% were found.

1 ICILS 2018 was conducted in the same year as the OECD's Teaching and Learning International Survey (TALIS) 2018 and Programme for International Student Assessment (PISA) 2018, and some national education surveys. The ICILS 2018 sampling team collaborated closely with the staff implementing sampling for these studies to prevent school sample overlap whenever possible. See Chapter 6 for further details.

Table B.3.2: Allocation of teacher sample in Finland

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools	Excluded schools
			Sampled	First replacement	Second replacement		
Helsinki/Uusimaa – Urban & semi-urban & rural	36	0	35	0	0	1	0
Southern – Urban & semi-urban	26	0	26	0	0	0	0
Southern – Rural	6	0	6	0	0	0	0
Western – Urban & semi-urban	28	1	24	0	0	2	1
Western – Rural	6	0	5	0	0	0	1
Northern & Eastern – Urban & semi-urban	28	1	26	1	0	0	0
Northern & Eastern – Rural	10	0	10	0	0	0	0
Swedish speaking – No Aland	8	0	8	0	0	0	0
Swedish speaking – Aland	2	0	2	0	0	0	0
Total	**150**	**2**	**142**	**1**	**0**	**3**	**2**

Note: One school with teacher participation rate below 50% was found.

B.4 France

- School level exclusions consisted of private schools without contract, schools in the overseas territories and in Mayotte, and specialized schools. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by school type (public school on priority education, public school priority education, private school), urbanization (less than 10,000 inhabitants, 10,000 to 200,000 inhabitants, more than 200,000 inhabitants), and equipment (large digital equipment, normal digital equipment), resulting in 18 explicit strata.
- No implicit stratification was applied.
- School sample overlap between ICILS 2018 and TALIS 2018: ICILS sample was selected using minimum overlap control to TALIS 2018.
- Schools with large digital equipment strata were oversampled to allow for better estimates.
- Small schools were selected with equal probabilities.

Table B.4.1: Allocation of student sample in France

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Public school – Non priority education – Less than 10,000 inhabitants – Large digital equipment	9	0	9	0	0	0
Public school – Non priority education – Less than 10,000 inhabitants – Normal digital equipment	24	0	24	0	0	0
Public school – Non priority education – 10,000 to 200,000 inhabitants – Large digital equipment	8	0	8	0	0	0
Public school – Non priority education – 10,000 to 200,000 inhabitants – Normal digital equipment	18	0	17	1	0	0
Public school – Non priority education – More than 200,000 inhabitants – Large digital equipment	12	0	12	0	0	0
Public school – Non priority education – More than 200,000 inhabitants – Normal digital equipment	19	0	19	0	0	0
Public school – Priority education – Less than 10,000 inhabitants – Large digital equipment	4	0	4	0	0	0
Public school – Priority education – Less than 10,000 inhabitants – Normal digital equipment	4	0	4	0	0	0
Public school – Priority education – 10,000 to 200,000 inhabitants – Large digital equipment	4	0	4	0	0	0
Public school – Priority education – 10,000 to 200,000 inhabitants – Normal digital equipment	6	0	6	0	0	0
Public school – Priority education More than 200,000 inhabitants – Large digital equipment	5	0	5	0	0	0
Public school – Priority education – More than 200,000 inhabitants – Normal digital equipment	8	0	8	0	0	0
Private school – Less than 10,000 inhabitants – Large digital equipment	4	0	4	0	0	0
Private school – Less than 10,000 inhabitants – Normal digital equipment	6	0	6	0	0	0
Private school – 10,000 to 200,000 inhabitants – Large digital equipment	4	0	4	0	0	0

Table B.4.1: Allocation of student sample in France (contd.)

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Private school – 10,000 to 200,000 inhabitants – normal digital equipment	8	0	8	0	0	0
Private school – More than 200,000 inhabitants – Large digital equipment	4	0	4	0	0	0
Private school – More than 200,000 inhabitants – Normal digital equipment	9	0	9	0	0	0
Total	**156**	**0**	**155**	**1**	**0**	**0**

Note: No schools with student participation rate below 50% were found.

Table B.4.2: Allocation of teacher sample in France

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Public school – Non priority education – Less than 10,000 inhabitants – Large digital equipment	9	0	6	0	0	3
Public school – Non priority education – Less than 10,000 inhabitants – Normal digital equipment	24	0	21	0	0	3
Public school – Non priority education – 10,000 to 200,000 inhabitants – Large digital equipment	8	0	8	0	0	0
Public school – Non priority education – 10,000 to 200,000 inhabitants – Normal digital equipment	18	0	15	0	0	3
Public school – Non priority education – More than 200,000 inhabitants – Large digital equipment	12	0	8	0	0	4
Public school – Non priority education – More than 200,000 inhabitants – Normal digital equipment	19	0	14	0	0	5
Public school – Priority education – Less than 10,000 inhabitants – Large digital equipment	4	0	3	0	0	1
Public school – Priority education – Less than 10,000 inhabitants – Normal digital equipment	4	0	3	0	0	1
Public school – Priority education – 10,000 to 200,000 inhabitants – Large digital equipment	4	0	2	0	0	2

Table B.4.2: Allocation of teacher sample in France (contd.)

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Public school – Priority education – 10,000 to 200,000 inhabitants – Normal digital equipment	6	0	4	0	0	2
Public school – Priority education More than 200,000 inhabitants – Large digital equipment	5	0	4	0	0	1
Public school – Priority education – More than 200,000 inhabitants – Normal digital equipment	8	0	3	0	0	5
Private school – Less than 10,000 inhabitants – Large digital equipment	4	0	4	0	0	0
Private school – Less than 10,000 inhabitants – Normal digital equipment	6	0	5	0	0	1
Private school – 10,000 to 200,000 inhabitants – Large digital equipment	4	0	4	0	0	0
Private school – 10,000 to 200,000 inhabitants – normal digital equipment	8	0	7	0	0	1
Private school – More than 200,000 inhabitants – Large digital equipment	4	0	3	0	0	1
Private school – More than 200,000 inhabitants – Normal digital equipment	9	0	8	0	0	1
Total	156	0	122	0	0	34

Note: Thirty-four schools were regarded as non-participating because the within-school participation rate was below 50%.

B.5 Germany

- School level exclusions consisted of special education schools and very small schools (less than three students in the target grade). Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification for regular schools was performed by federal state (North Rhine-Westphalia, other federal states) and school type (Gymnasium, non-Gymnasium). Special education schools with students able to do the test were placed in a separate stratum, resulting in five explicit strata.
- Implicit stratification for regular schools was applied by socioeconomic status predictor (3 levels) and federal state (16 levels). For special education schools, implicit stratification was done by federal states (16 levels), giving a total of 65 implicit strata.
- School sample overlap between ICILS 2018, PISA 2018, and national Assessment Educational Standards 2018 (Bildungstrend 2018): ICILS sample was selected using minimum overlap control to both surveys.
- The sample was disproportionally allocated to explicit strata.
- Schools in North Rhine-Westphalia were oversampled due to the benchmark characteristic.
- Small schools were selected with equal probabilities.

Table B.5.1: Allocation of student sample in Germany

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools	Excluded schools
			Sampled	First replacement	Second replacement		
North Rhine-Westphalia – Gymnasium	42	0	38	4	0	0	0
North Rhine-Westphalia – Non-Gymnasium	72	3	65	1	0	3	0
Other federal states – Gymnasium	44	0	37	1	2	4	0
Other federal states – Non-Gymnasium	72	0	51	4	3	13	1
Special education schools – None	4	1	2	1	0	0	0
Total	**234**	**4**	**193**	**11**	**5**	**20**	**1**

Note: Six schools were regarded as non-participating because the within-school participation rate was below 50%.

Table B.5.2: Allocation of teacher sample in Germany

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools	Excluded schools
			Sampled	First replacement	Second replacement		
North Rhine-Westphalia – Gymnasium	42	0	36	4	0	2	0
North Rhine-Westphalia – Non-Gymnasium	72	3	65	1	0	3	0
Other federal states – Gymnasium	44	0	25	1	1	17	0
Other federal states – Non-Gymnasium	72	0	41	4	2	24	1
Special education schools – None	4	1	2	0	0	1	0
Total	**234**	**4**	**169**	**10**	**3**	**47**	**1**

Note: Thirty-three schools were regarded as non-participating because the within-school participation rate was below 50%.

B.6 Italy

- School level exclusions consisted of schools for children with special needs, schools with less than six students in the target grade, schools with Slovenian instruction language, and schools in remote areas or on little islands. Within-school exclusions consisted of functionally disabled students.
- Explicit stratification was performed by geographic region (North, Central, South).
- Implicit stratification was applied by school type (public, private) and performance group (5), giving a total of 30 implicit strata.
- Small schools were selected with equal probabilities.

Table B.6.1: Allocation of student sample in Italy

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
North	66	0	66	0	0	0
Central	28	0	24	4	0	0
South	56	0	53	3	0	0
Total	**150**	**0**	**143**	**7**	**0**	**0**

Note: No schools with student participation rate below 50% were found.

Table B.6.2: Allocation of teacher sample in Italy

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
North	66	0	66	0	0	0
Central	28	0	24	4	0	0
South	56	0	51	3	0	2
Total	**150**	**0**	**141**	**7**	**0**	**2**

Note: Two schools were regarded as non-participating because the within-school participation rate was below 50%.

B.7 Kazakhstan

- School level exclusions consisted of schools for children with special needs, schools with less than five students in the target grade, Uighur schools, Uzbek schools, Tadjik schools, and other language schools. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by urbanization (urban, rural) and language of instruction (4), resulting in eight explicit strata.
- No implicit stratification was applied.
- The sample was disproportionally allocated to explicit strata.
- Schools were oversampled to allow for better estimates for the different language groups.
- Small schools were selected with equal probabilities.

Table B.7.1: Allocation of student sample in Kazakhstan

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Urban – Kazakh only	30	0	30	0	0	0
Urban – Russian only	20	0	20	0	0	0
Urban – Kazakh & Russian	30	0	30	0	0	0
Urban – Other	14	1	13	0	0	0
Rural – Kazakh only	36	0	36	0	0	0
Rural – Russian only	10	0	10	0	0	0
Rural – Kazakh & Russian	30	0	29	0	0	1
Rural – Other	16	1	15	0	0	0
Total	**186**	**2**	**183**	**0**	**0**	**1**

Note: One school with student participation rate below 50% was found.

Table B.7.2: Allocation of teacher sample in Kazakhstan

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Urban – Kazakh only	30	0	30	0	0	0
Urban – Russian only	20	0	20	0	0	0
Urban – Kazakh & Russian	30	0	30	0	0	0
Urban – Other	14	1	13	0	0	0
Rural – Kazakh only	36	0	36	0	0	0
Rural – Russian only	10	0	10	0	0	0
Rural – Kazakh & Russian	30	0	30	0	0	0
Rural – Other	16	1	15	0	0	0
Total	**186**	**2**	**184**	**0**	**0**	**0**

Note: No schools with teacher participation rate below 50% were found.

B.8 Korea, Republic of

- School level exclusions consisted of geographically inaccessible schools, schools with less than five students in the target grade, and schools with different curriculum (physical education schools). Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by urbanization (urban, suburban, rural) and school type (boys, girls, mixed), resulting in nine explicit strata.
- No implicit stratification was applied.
- The sample was disproportionally allocated to explicit strata.
- Schools were oversampled to allow for better estimates for rural, boys, and girls schools.
- Small schools were selected with equal probabilities.

Table B.8.1: Allocation of student sample in Korea

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Urban – Boys	10	0	10	0	0	0
Urban – Girls	10	0	10	0	0	0
Urban – Mixed	34	0	34	0	0	0
Suburban – Boys	12	0	12	0	0	0
Suburban – Girls	12	0	12	0	0	0
Suburban – Mixed	38	0	38	0	0	0
Rural – Boys	8	0	8	0	0	0
Rural – Girls	8	0	8	0	0	0
Rural – Mixed	18	0	18	0	0	0
Total	**150**	**0**	**150**	**0**	**0**	**0**

Note: No schools with student participation rate below 50% were found.

Table B.8.2: Allocation of teacher sample in Korea

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Urban – Boys	10	0	10	0	0	0
Urban – Girls	10	0	10	0	0	0
Urban – Mixed	34	0	34	0	0	0
Suburban – Boys	12	0	11	0	0	1
Suburban – Girls	12	0	12	0	0	0
Suburban – Mixed	38	0	36	0	0	2
Rural – Boys	8	0	8	0	0	0
Rural – Girls	8	0	8	0	0	0
Rural – Mixed	18	0	18	0	0	0
Total	**150**	**0**	**147**	**0**	**0**	**3**

Note: No schools with teacher participation rate below 50% were found.

B.9 Luxembourg

- No school-level exclusions. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by curriculum (school following national curriculum, schools following different curriculum), resulting in two explicit strata.
- No implicit stratification was applied.
- Census of all schools and students. All variance estimates were computed using schools as variance strata.

Table B.9.1: Allocation of student sample in Luxembourg

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Schools following national curriculum	33	0	32	0	0	1
Schools with different curriculum	8	0	6	0	0	2
Total	**41**	**0**	**38**	**0**	**0**	**3**

Note: No schools with student participation rate below 50% were found.

Table B.9.2: Allocation of teacher sample in Luxembourg

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Schools following national curriculum	33	0	23	0	0	10
Schools with different curriculum	8	0	5	0	0	3
Total	**41**	**0**	**28**	**0**	**0**	**13**

Note: Ten schools were regarded a non-participating because the within-school participation rate was below 50%.

B.10 Portugal

- School-level exclusions consisted of schools with less than seven students in the target grade, and international schools. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was implemented by school type (public, private) and subregions (23), resulting in 28 explicit strata.
- No implicit stratification was applied.
- Small school samples within regions necessitated disproportional sample allocations.
- Small schools were selected with equal probabilities.

Table B.10.1: Allocation of student sample in Portugal

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Public – Alto Minho	6	0	6	0	0	0
Public – Cávado	6	0	6	0	0	0
Public – Ave	6	0	6	0	0	0
Public – Área Metropolitana do Porto	18	0	17	0	0	1
Public – Alto Tâmega	6	0	6	0	0	0
Public – Tâmega e Sousa	6	0	6	0	0	0
Public – Douro	6	0	6	0	0	0
Public – Terras de Trás-os-Montes	6	0	6	0	0	0
Public – Oeste	6	0	5	0	0	1
Public – Região de Aveiro	6	0	6	0	0	0
Public – Região de Coimbra	6	0	5	1	0	0
Public – Região de Leiria	6	0	5	0	0	1
Public – Viseu Dão Lafões	6	0	6	0	0	0
Public – Beira Baixa	6	0	6	0	0	0
Public – Médio Tejo	6	0	5	1	0	0
Public – Beiras e Serra da Estrela	6	0	5	1	0	0
Public – Área Metropolitana de Lisboa	28	0	23	0	0	5
Public – Alentejo Litoral	6	0	4	0	0	2
Public – Baixo Alentejo	6	0	6	0	0	0
Public – Lezíria do Tejo	6	0	6	0	0	0
Public – Alto Alentejo	6	0	6	0	0	0
Public – Alentejo Central	6	0	4	0	0	2
Public – Algarve	6	0	5	0	0	1
Public – Região Autónoma dos Açores	6	0	2	1	0	3
Public – Região Autónoma da Madeira	6	0	6	0	0	0
Private – Área Metropolitana do Porto	7	0	5	2	0	0
Private – Área Metropolitana de Lisboa	7	0	5	1	0	1
Private – Other Regions	22	1	15	4	0	2
Total	**220**	**1**	**189**	**11**	**0**	**19**

Note: Seventeen schools were regarded as non-participating because the within-school participation rate was below 50%.

Table B.10.2: Allocation of teacher sample in Portugal

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Public - Alto Minho	6	0	6	0	0	0
Public – Cávado	6	0	6	0	0	0
Public – Ave	6	0	6	0	0	0
Public – Área Metropolitana do Porto	18	0	17	0	0	1
Public – Alto Tâmega	6	0	6	0	0	0
Public – Tâmega e Sousa	6	0	5	0	0	1
Public – Douro	6	0	6	0	0	0
Public – Terras de Trás-os-Montes	6	0	6	0	0	0
Public – Oeste	6	0	5	0	0	1
Public – Região de Aveiro	6	0	6	0	0	0
Public – Região de Coimbra	6	0	5	1	0	0
Public – Região de Leiria	6	0	6	0	0	0
Public – Viseu Dão Lafões	6	0	6	0	0	0
Public – Beira Baixa	6	0	6	0	0	0
Public – Médio Tejo	6	0	5	1	0	0
Public – Beiras e Serra da Estrela	6	0	5	1	0	0
Public – Área Metropolitana de Lisboa	28	0	26	0	0	2
Public – Alentejo Litoral	6	0	5	0	0	1
Public – Baixo Alentejo	6	0	6	0	0	0
Public – Lezíria do Tejo	6	0	6	0	0	0
Public – Alto Alentejo	6	0	6	0	0	0
Public – Alentejo Central	6	0	6	0	0	0
Public – Algarve	6	0	6	0	0	0
Public – Região Autónoma dos Açores	6	0	4	1	0	1
Public – Região Autónoma da Madeira	6	0	6	0	0	0
Private – Área Metropolitana do Porto	7	0	5	2	0	0
Private – Área Metropolitana de Lisboa	7	0	6	1	0	0
Private – Other Regions	22	1	13	4	0	4
Total	**220**	**1**	**197**	**11**	**0**	**11**

Note: Seven schools were regarded as non-participating because the within-school participation rate was below 50%.

B.11 United States

- School-level exclusions consisted of schools with less than two classes in the target grade and private schools. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by poverty level (2), school type (public, private), and geographic regions (Northeast, Midwest, South, West), resulting in 12 explicit strata.
- Implicit stratification was applied by school location (city, rural, suburban, town), and ethnicity status, giving a total of 96 implicit strata.
- Small schools were selected with equal probabilities.

Table B.11.1: Allocation of student sample in the United States

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools	Excluded schools
			Sampled	First replacement	Second replacement		
High poverty – Public – Northeast	17	0	6	1	0	10	0
High poverty – Public – Midwest	25	1	18	0	0	6	0
High poverty – Public – South	68	1	58	3	2	4	0
High poverty – Public – West	42	1	35	4	0	2	0
Low poverty – Private – Northeast	7	0	1	2	0	4	0
Low poverty -– Private – Midwest	7	0	4	1	0	2	0
Low poverty – Private – South	10	1	5	0	0	4	0
Low poverty – Private – West	6	1	1	0	1	3	0
Low poverty – Public – Northeast	34	0	12	2	1	19	0
Low poverty – Public – Midwest	43	1	24	5	2	11	0
Low poverty – Public – South	56	1	40	5	1	8	1
Low poverty – Public – West	37	1	27	2	0	6	1
Total	**352**	**8**	**231**	**25**	**7**	**79**	**2**

Note: Three schools were regarded as non-participating because the within-school participation rate was below 50%.

Table B.11.2: Allocation of teacher sample in the United States

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools	Excluded schools
			Sampled	First replacement	Second replacement		
High poverty – Public – Northeast	17	0	5	1	0	11	0
High poverty – Public – Midwest	25	1	17	0	0	7	0
High poverty – Public – South	68	1	57	3	2	5	0
High poverty – Public – West	42	1	33	4	0	4	0
Low poverty – Private – Northeast	7	0	1	2	1	3	0
Low poverty -- Private – Midwest	7	0	4	1	0	2	0
Low poverty – Private – South	10	1	5	0	0	4	0
Low poverty – Private – West	6	1	1	0	1	3	0
Low poverty – Public – Northeast	34	0	12	2	1	19	0
Low poverty – Public – Midwest	43	1	25	5	2	10	0
Low poverty – Public – South	56	1	40	5	1	8	1
Low poverty – Public – West	37	1	26	2	0	7	1
Total	**352**	**8**	**226**	**25**	**8**	**83**	**2**

Note: Seven schools were regarded as non-participating because the within-school participation rate was below 50%.

B.12 Uruguay

- School-level exclusions consisted of schools for children with special needs and rural schools. Within-school exclusions consisted of functionally disabled students.
- Explicit stratification was performed by school type (public, private), location (Montevideo, Interior Urban), and school type (high school/Lyceum, vocational school/Utu) resulting in six explicit strata.
- No implicit stratification was applied.
- The sample was disproportionally allocated to explicit strata. Schools were oversampled to allow for better estimates for public and private schools, Montevideo and Interior Urban schools, as well as Lyceum (high school) and Utu schools (vocational schools).
- To enable overlap control for PISA 2018, schools in one stratum were selected with equal probabilities, and in two other strata the selection probabilities have been capped to 0.5.
- Within census strata all variance estimates were computed using schools as variance strata.
- Small schools were selected with equal probabilities.

Table B.12.1: Allocation of student sample in Uruguay

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Public – Montevideo – Lyceum	30	1	23	3	0	3
Public – Montevideo – Utu	27	1	25	0	0	1
Public – Interior urban – Lyceum	60	1	57	1	0	1
Public – Interior urban – Utu	30	1	28	0	0	1
Private – Montevideo – Lyceum	18	0	15	3	0	0
Private – Interior urban – Lyceum	12	1	11	0	0	0
Total	**177**	**5**	**159**	**7**	**0**	**6**

Note: Six schools were regarded as non-participating because the within-school participation rate was below 50%.

Table B.12.2: Allocation of teacher sample in Uruguay

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Public – Montevideo – Lyceum	30	1	18	2	0	9
Public – Montevideo – Utu	27	1	11	0	0	15
Public – Interior urban – Lyceum	60	1	42	1	0	16
Public – Interior urban – Utu	30	1	24	0	0	5
Private – Montevideo – Lyceum	18	0	12	3	0	3
Private – Interior urban – Lyceum	12	1	8	0	0	3
Total	**177**	**5**	**115**	**6**	**0**	**51**

Note: Fifty-one schools were regarded as non-participating because the within-school participation rate was below 50%.

Benchmarking participants

B.13 Moscow, Russian Federation

- School-level exclusions consisted of schools with less than seven students. Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification was performed by school performance.
- Implicit stratification was applied by school type (public, private), giving a total of 27 implicit strata.
- In first two explicit strata schools were selected with equal probabilities.
- Small schools were selected with equal probabilities.

Table B.13.1: Allocation of student and teacher sample in Moscow, Russian Federation

School participation status—Student and teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Top 1-50 schools	26	0	26	0	0	0
Top 51-100 schools	26	0	26	0	0	0
Top 101-200 schools	30	0	30	0	0	0
Top 201-300 schools	30	0	30	0	0	0
Top 301 and above-all other schools	38	0	36	1	1	0
Total	**150**	**0**	**148**	**1**	**1**	**0**

Note: No schools with student participation rate below 50% were found. No schools with teacher participation rate below 50% were found.

B.14 North Rhine-Westphalia, Germany

- School-level exclusions consisted of special education schools and very small schools (less than three students in the target grade). Within-school exclusions consisted of intellectually disabled students, functionally disabled students, and non-native language speakers.
- Explicit stratification for regular schools was performed by school type (Gymnasium, non-Gymnasium). Special education schools with students able to do the test were placed in a separate stratum, resulting in three explicit strata.
- Implicit stratification for regular schools was applied by socioeconomic status predictor (3 levels), giving a total of four implicit strata.
- School sample overlap between ICILS 2018, PISA 2018, and national Assessment Educational Standards 2018: ICILS sample was selected using minimum overlap control to both surveys.
- Small schools were selected with equal probabilities.

Table B.14.1: Allocation of student sample in North Rhine-Westphalia, Germany

School participation status—Student survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Gymnasium	42	0	38	4	0	0
Non-Gymnasium	72	3	65	1	0	3
Special education schools	1	0	1	0	0	0
Total	**115**	**3**	**104**	**5**	**0**	**3**

Note: No schools with student participation rate below 50% were found.

Table B.14.2: Allocation of teacher sample in North Rhine-Westphalia, Germany

School participation status—Teacher survey

Explicit strata	Total sampled schools	Ineligible schools	Participating schools			Non-participating schools
			Sampled	First replacement	Second replacement	
Gymnasium	42	0	36	4	0	2
Non-Gymnasium	72	3	65	1	0	3
Special education schools	1	0	1	0	0	0
Total	**115**	**3**	**102**	**5**	**0**	**5**

Note: Three schools were regarded as non-participating because the within-school participation rate was below 50%.

APPENDIX C:

National items excluded from scaling

Country	Set to "not administered" in database	Reason for deletion	Excluded from scaling	Reason for exclusion
Chile			B09C, R12B	B09C, R12BB09C: country DIF > 1.3 R12B: country DIF > 1.3
Denmark			R12C	R12C: country DIF > 1.3
Finland			S03Z	S03Z: country DIF > 1.3
France			H07B, H07C, H07D, H07E, H07F, S08F, G03Z, G09C, G09G, R06A, R12C, R12D	H07B: Inter Coder Reliability < 70% H07C: Inter Coder Reliability < 70% H07D: Inter Coder Reliability < 70% H07E: Inter Coder Reliability < 70% H07F: Inter Coder Reliability < 70% S08F: Inter Coder Reliability < 70% G03Z: Inter Coder Reliability < 70% G09C: Inter Coder Reliability < 70% G09G: Inter Coder Reliability < 70% R06A: Inter Coder Reliability < 70% R12C: Inter Coder Reliability < 70% R12D: Inter Coder Reliability < 70%
Germany			R12C, H07C, H07D, H07F, G03Z, R12A	R12C: country DIF > 1.3 H07C: Inter Coder Reliability < 70% H07D: Inter Coder Reliability < 70% H07F: Inter Coder Reliability < 70% G03Z: Inter Coder Reliability < 70% R12A: Inter Coder Reliability < 70%
Italy			H07E, S06Z	H07E: country DIF > 1.3 S06Z: country DIF > 1.3
Kazakhstan			S04B, R12C	S04B: country DIF > 1.3 R12C: country DIF > 1.3
Korea, Republic of	B09E		B02Z, G06Z, G08Z, G09C, R12B, R12C, R12J	B02Z: country DIF > 1.3 G06Z: country DIF > 1.3 G08Z: country DIF > 1.3 G09C: country DIF > 1.3 R12B: country DIF > 1.3 R12C: country DIF > 1.3 R12J: country DIF > 1.3
Moscow (Russian Federation)			H07J, G03Z	H07J: country DIF > 1.3 G03Z: country DIF > 1.3

Country	Set to "not administered" in database	Reason for deletion	Excluded from scaling	Reason for exclusion
North Rhine-Westphalia (Germany)			R12C	R12C: country DIF > 1.3
Portugal			H07J, G08Z, H07B, H07F, H07G	H07J: country DIF > 1.3 G08Z: country DIF > 1.3 H07B: Inter Coder Reliability < 70% H07F: Inter Coder Reliability < 70% H07G: Inter Coder Reliability < 70%
United States			B09D, H07F, G07Z, R07Z, R12C, H07E	B09D: country DIF > 1.3 H07F: country DIF > 1.3 G07Z: country DIF > 1.3 R07Z: country DIF > 1.3 R12C: country DIF > 1.3 H07E: Inter Coder Reliability < 70%
Uruguay			H07J, G03Z, R06A, H07B, H07C, H07D, H07F, H07G, H07H, H07I, G08Z, G09C, G09D, G09G, R08Z, R10Z	H07J: country DIF > 1.3 G03Z: country DIF > 1.3 and Inter Coder Reliability < 70% R06A: country DIF > 1.3 and Inter Coder Reliability < 70% H07B: Inter Coder Reliability < 70% H07C: Inter Coder Reliability < 70% H07D: Inter Coder Reliability < 70% H07F: Inter Coder Reliability < 70% H07G: Inter Coder Reliability < 70% H07H: Inter Coder Reliability < 70% H07I: Inter Coder Reliability < 70% G08Z: Inter Coder Reliability < 70% G09C: Inter Coder Reliability < 70% G09D: Inter Coder Reliability < 70% G09G: Inter Coder Reliability < 70% R08Z: Inter Coder Reliability < 70% R10Z: Inter Coder Reliability < 70%

APPENDIX D:

Student background variables used for conditioning

Variable	Name	Values	Coding	Regressor
Stratum ID	IDSTRATE_[X]		Dummy variable per stratum with the largest stratum as reference category	Direct
Adjusted school mean achievement	SCH_MN	Logits	Value	Direct
Age	AGE	Value Missing	Copy,0 Mean,1	PCA
Gender	GENDER	Female Male Missing	10 00 01	Direct
Highest level of parental education	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] Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Highest level of parental occupation	HISEI	Value Missing	Copy,0 Mean,1	PCA
Student expected levels of education to complete	SISCED	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] Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Country of birth – You Country of birth – Mother or [female guardian] Country of birth – Father or [male guardian]	IS2G04A IS2G04B IS2G04C	1 = [Country of test] 2 = [Other country/Group A] 3 = [Other country/Group B] 4 = [Another country] Missing	Length of dummy code varies across countries, option with the highest percentage is reference category (Missing = reference category)	PCA
Language at home	IS2G05	1 = [Language of test] 2 = [Other language 1] 3 = [Other language 2] 4 = [Another language] Missing	Length of dummy code varies across countries, option with the highest percentage is reference category (Missing = reference category)	PCA
Books at home	IS2G14	1. None or very few (0–10 books) 2. Enough to fill one shelf (11–25 books) 3. Enough to fill one bookcase (26–100 books) 4. Enough to fill two bookcases (101–200 books) 5. Enough to fill three or more bookcases more than 200 books) Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA

Variable	Name	Values	Coding	Regressor
Devices at home – Desktop/laptop Devices at home – Tablet/e-reader	IS2G15AA IS2G15AB	1. None 2. One 3. Two 4. Three or more Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Internet connection at home	IS2G15B	1. Yes 2. No Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
How long using desktop/laptop computers How long using tablets/e-readers How long using smartphones	IS2G16A IS2G16B IS2G16C	1. Less than one year 2. At least one year but less than three years 3. At least three years but less than five years 4. At least five years but less than seven years 5. Seven years or more Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Teaching - Communicating over the Internet Teaching - Edit/creating digital documents Teaching - Edit/creating digital presentations Teaching - Changing settings on ICT device Teaching - Finding information on the Internet Teaching - Working in a computer network	IS2G17A IS2G17B IS2G17C IS2G17D IS2G17E IS2G17F	1. My teachers 2. My family 3. My friends 4. I taught myself 5. I have never learned this Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
How often using ICT – At school for school-related purposes How often using ICT – At school for other purposes How often using ICT – Outside school for school-related purposes How often using ICT – Outside school for other purposes	IS2G18A IS2G18B IS2G18C IS2G18D	1. Never 2. Less than once a month 3. At least once a month but not every week 4. At least once a week but not every day 5. Every day Missing 4. At least once a week but not every day	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA

Variable	Name	Values	Coding	Regressor
Using ICT – Write or edit documents	IS2G19A	1. Never 2. Less than once a month 3. At least once a month but not every week 4. At least once a week but not every day 5. Every day Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Using ICT – Use a spreadsheet to do calculations, store data or plot graphs (e.g. using [Microsoft Excel ®])	IS2G19B			
Using ICT – Create a simple "slideshow" presentation (e.g. using [Microsoft PowerPoint ®])	IS2G19C			
Using ICT – Record or edit videos	IS2G19D			
Using ICT – Write computer programs, scripts or apps (e.g. using [Logo, LUA or Scratch])	IS2G19E			
Using ICT – Use drawing, painting or graphics software or [apps]	IS2G19F			
Using ICT – Produce or edit music	IS2G19G			
Using ICT – Build a webpage	IS2G19H			
Using ICT – Share news about current events on social media	IS2G20A	1. Never 2. Less than once a month 3. At least once a month but not every week 4. At least once a week but not every day 5. Every day Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Using ICT – Communicate with friends, family, or other people using instant messaging, voice or video chat (e.g. [Skype, WhatsApp, Viber])	IS2G20B			
Using ICT – Send texts or instant messages to friends, family, or other people	IS2G20C			
Using ICT – Write posts and updates about what happens in your life on social media	IS2G20D			
Using ICT – Ask questions on forums or [Q&A, question and answer] websites	IS2G20E			
Using ICT – Answer other peoples' questions on forums or [Q&A, question and answer] websites	IS2G20F			
Using ICT – Write posts for your own blog (e.g. [WordPress, Tumblr, Blogger])	IS2G20G			
Using ICT – Post images or video in social networks or online communities (e.g. [Facebook, Instagram or YouTube])	IS2G20H			
Using ICT – Watch videos or images that other people have posted online	IS2G20I			
Using ICT – Send or forward information about events or activities to other people	IS2G20J			

Variable	Name	Values	Coding	Regressor
Using ICT – Search the Internet to find information about places to go or activities to do Using ICT – Read reviews on the Internet of things you might want to buy Using ICT – Read news stories on the Internet Using ICT – Search for online information about things you are interested in Using ICT – Use websites, forums, or online videos o find out how to do something Using ICT – Play single-player games Using ICT – Listen to downloaded or streamed music Using ICT – Watch downloaded or streamed TV shows or movies	IS2G21A IS2G21B IS2G21C IS2G21D IS2G21E IS2G21F IS2G21G IS2G21H	1. Never 2. Less than once a month 3. At least once a month but not every week 4. At least once a week but not every day 5. Every day 4. At least once a week but not every day Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
ICT use – Prepare reports or essays ICT use – Prepare presentations ICT use – Work online with other students ICT use – Complete [worksheets] or exercises ICT use – Organise your time and work ICT use – Take tests ICT use – Use software or applications to learn skills or a subject (e.g. mathematics tutoring software, language learning software) ICT use – Use the Internet to do research ICT use – Use coding software to complete assignments (e.g. [Scratch]) ICT use – Make video or audio productions	IS2G22A IS2G22B IS2G22C IS2G22D IS2G22E IS2G22F IS2G22G IS2G22H IS2G22I IS2G22J	1. Never 2. Less than once a month 3. At least once a month but not every week 4. At least once a week but not every day 5. Every day Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
ICT use – [Language arts: test language] ICT use – [Language arts: foreign and other national languages] ICT use – Mathematics [Add any appropriate national examples] ICT use – Sciences (general science and/or physics, chemistry, biology, geology, earth sciences) ICT use – Human sciences /Humanities / Social studies (history, geography, civics, law, economics, etc.) ICT use – Creative arts ([visual] arts, music, dance, drama, etc.) ICT use – [Information technology, computer studies or similar] ICT use – Practical or vocational [Add any appropriate national examples] ICT use – Other (e.g. [moral/ethics, physical education, personal and social development])	IS2G23A IS2G23B IS2G23C IS2G23D IS2G23E IS2G23F IS2G23G IS2G23H IS2G23I	1. I don't study this subject / these subjects 2. Never 3. In some lessons 4. In most lesson 5. In every or almost every lesson Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA

Variable	Name	Values	Coding	Regressor
Tool use – Tutorial software or [practice programs]	IS2G24A	1. Never 2. In some lessons 3. In most lessons 4. In every or almost every lesson Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Tool use – Word-processing software (e.g. [Microsoft Word ®])	IS2G24B			
Tool use – Presentation software (e.g. [Microsoft PowerPoint ®])	IS2G24C			
Tool use – Spreadsheets (e.g. [Microsoft Excel®])	IS2G24D			
Tool use – Multimedia production tools (e.g. media capture and editing, web production)	IS2G24E			
Tool use – Concept mapping software (e.g. [Inspiration ®], [Webspiration ®])	IS2G24F			
Tool use – Tools that capture real-world data (e.g. speed, temperature) digitally for analysis	IS2G24G			
Tool use – Simulations and modelling software	IS2G24H			
Tool use – Computer-based information resources (e.g. websites, wikis, encyclopaedia)	IS2G24I			
Tool use – Interactive digital learning resources (e.g. learning games or applications)	IS2G24J			
Tool use – Graphing or drawing software	IS2G24K			
School learning – Provide references to Internet sources	S2G25A	1. To a large extent 2. To a moderate extent 3. To a small extent 4. Not at all Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
School learning – Search for information using ICT	IS2G25B			
School learning – Present information for a given audience or purpose using ICT	IS2G25C			
School learning – Work out whether to trust information from the Internet	IS2G25D			
School learning – Decide what information obtained from the Internet is relevant to include in school work	IS2G25E			
School learning – Organize information obtained from Internet sources	IS2G25F			
School learning – Decide where to look for information on the Internet about an unfamiliar topic	IS2G25G			
School learning – Use ICT to collaborate with others	IS2G25H			

Variable	Name	Values	Coding	Regressor
School learning – To change passwords regularly (e.g. network account, email, social media) School learning – To check the origin of emails before opening attachments School learning – To log out of a shared computer at the end of a session School learning – To share information on social media responsibly	IS2G26A IS2G26B IS2G26C IS2G26D	1. Yes 2. No Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Do well – Edit digital photographs or other graphic images Do well – Create a database (e.g. using [Microsoft Access ®]) Do well – Write or edit text for a school assignment Do well – Search for and find relevant information for a school project on the Internet Do well – Build or edit a webpage Do well – Change the settings on your device to improve the way it operates Do well – Create a computer program, macro, or [app] (e.g. in [Basic, Visual Basic]) Do well – Set up a local area network of computers or other ICT Do well – Create a multi-media presentation (with sound, pictures, or video) Do well – Upload text, images, or video to an online profile Do well – Insert an image into a document or message Do well – Install a program or [app] Do well – Judge whether you can trust information you find on the Internet	IS2G27A IS2G27B IS2G27C IS2G27D IS2G27E IS2G27F IS2G27G IS2G27H IS2G27I IS2G27J IS2G27K IS2G27L IS2G27M	1. I know how to do this 2. I have never done this but I could work out how to do this 3. I do not think I could do this Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA

Variable	Name	Values	Coding	Regressor
Advances in ICT usually improve people's living conditions. ICT helps us to understand the world better. Using ICT makes people more isolated in society. With more ICT there will be fewer jobs. People spend far too much time using ICT. ICT is valuable to society. Advances in ICT bring many social benefits. Using ICT may be dangerous for people's health. I would like to study subjects related to ICT after [secondary school]. I hope to find a job that involves advanced ICT. Learning how to use ICT applications will help me to do the work I am interested in.	IS2G28A IS2G28B IS2G28C IS2G28D IS2G28E IS2G28F IS2G28G IS2G28H IS2G28I IS2G28J IS2G28K	1. Strongly agree 2. Agree 3. Disagree 4. Strongly disagree Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Coding tasks at school – To display information in different ways Coding tasks at school – To break a complex process into smaller parts Coding tasks at school – To understand diagrams that describe or show real-world problems Coding tasks at school – To plan tasks by setting out the steps needed to complete them Coding tasks at school – To use tools to make diagrams that help solve problems Coding tasks at school – To use simulations to help understand or solve real world problems Coding tasks at school – To make flow diagrams to show the different parts of a process Coding tasks at school – To record and evaluate data to understand and solve a problem Coding tasks at school – To use real-world data to review and revise solutions to problems	IS2G29A IS2G29B S2G29C IS2G29D IS2G29E IS2G29F IS2G29G IS2G29H IS2G29I	1. Strongly agree 2. Agree 3. Disagree 4. Strongly disagree Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA
Study [computing, computer science, information technology, informatis or similar] in the current school year	IS2G30	1. Yes 2. No Missing	Dummy coding, option with highest percentage is reference category (Missing = reference category)	PCA

IEA's International Computer and Information Literacy Study (ICILS) 2018 investigated how well students are prepared for study, work, and life in a digital world. ICILS 2018 measured international differences in students' computer and information literacy (CIL): their ability to use computers to investigate, create, participate, and communicate at home, at school, in the workplace, and in the community. Participating countries had an additional option for their students to complete an assessment of computational thinking (CT): their ability to recognize aspects of real-world problems appropriate for computational formulation, and to evaluate and develop algorithmic solutions to those problems, so that the solutions could be operationalized with a computer.

This technical report follows the publication of several international and regional reports that presented the results of ICILS 2018. It provides a comprehensive account of the conceptual, methodological, and analytical implementation of the study. It includes detailed information on the development of the data-collection instruments used, including their translation and translation verification, on sampling design and implementation, sampling weights and participation rates, survey operation procedures, quality control of data collection, data management and creation of the international database, scaling procedures, and analysis of ICILS 2018 data. The technical report enables researchers to evaluate published reports and articles based on data from this study and, used in conjunction with the ICILS 2018 User Guide for the International Database, will provide guidance for their own analyses.