International Computer and Information Literacy Study

ICILS 2013 Technical Report

Edited by
Julian Fraillon
Wolfram Schulz
Tim Friedman
John Ainley
Eveline Gebhardt

International Computer and Information Literacy Study

ICILS 2013 Technical Report

Edited by
Julian Fraillon
Wolfram Schulz
Tim Friedman
John Ainley
Eveline Gebhardt

Contributors
John Ainley, Ralph Carstens, Diego Cortes, David Ebbs, Julian Fraillon, Tim Friedman, Eveline Gebhardt, Michael Jung, Paulína Koršňáková, Sabine Meinck, Wolfram Schulz

ISBN: 978-90-79549-30-6
DOI: 10.15478/uuid:b9cdd888-6665-4e9f-a21e-61569845ed5b

Publisher: IEA Secretariat
Herengracht 487
1017 BT Amsterdam, the Netherlands

Telephone: +31 20 625 3625
Fax: + 31 20 420 7136
Email: Department@IEA.nl
Website: www.iea.nl

The International Association for the Evaluation of Educational Achievement, known as IEA, is an independent, international consortium of national research institutions and governmental research agencies, with headquarters in Amsterdam. Its primary purpose is to conduct large-scale comparative studies of educational achievement with the aim of gaining more in-depth understanding of the effects of policies and practices within and across systems of education.

Copyedited by Paula Wagemaker Editorial Services, Oturehua, Central Otago, New Zealand
Design and production by Becky Bliss Design and Production, Wellington, New Zealand

Contents

List of Tables and Figures

Tables

Figures

CHAPTER 1:

Overview of ICILS

John Ainley and Julian Fraillon

Introduction

The International Computer and Information Literacy Study (ICILS) studied the extent to which young people have developed computer and information literacy (CIL) to support their capacity to participate in the digital age. Many countries recognize the importance that education in information and communication technologies (ICT) has for enabling citizens to develop the competencies needed to access information and participate in transactions using ICT (European Commission, 2006). We refer to these competencies as *computer and information literacy*, which is defined as "an individual's ability to use computers to investigate, create and communicate in order to participate effectively at home, at school, in the workplace and in society" (Fraillon, Schulz, & Ainley, 2013, p. 17).

ICILS systematically investigated differences in CIL outcomes across the participating countries and looked at how these countries provide CIL-related education. The study also explored differences within and across countries with respect to the relationship between the outcomes of CIL education and student characteristics and school contexts.

ICILS based its investigation on four research questions concerned with:

1. Variations in CIL between and within countries;
2. Aspects of schools, education systems, and teaching associated with student achievement in CIL;
3. The extent to which students' access to, familiarity with, and self-reported proficiency in using computers is associated with student achievement in CIL; and
4. The aspects of students' personal and social backgrounds that are associated with CIL.

The ICILS assessment framework (Fraillon et al., 2013) describes the development of these questions. The framework also provides more details relating to the questions and outlines the variables necessary for analyses associated with the questions.

Instruments

CIL test

The student assessment was based on four 30-minute modules. Each of the four assessment modules consisted of a set of questions and tasks based on a realistic theme and following a linear narrative structure. A series of small discrete tasks (typically taking less than a minute to complete) preceded a large task that typically took 15 to 20 minutes to complete. Collectively, the modules contained a total of 62 tasks and questions corresponding to 82 score points.

Each test completed by a student consisted of two of the four modules, so the overall assessment time for each student was one hour. In total, there were 12 different possible combinations of module pairs. Each module appeared in six of the combinations—three times as the first and three times as the second module when paired with each

of the other three. 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 assessment framework. The design also controlled for the influence of item position on difficulty across the sampled students and provided a variety of contexts for the assessment of CIL.

When students began each module, they were first presented with an overview of the theme and purpose of the tasks in the module. The overview also included a basic description of what the large task would comprise. Each module was designed as a narrative, which meant that the module's smaller discrete tasks required the students to complete 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 four modules were:

- *After School Exercise:* Students set up an online collaborative workspace to share information and then selected and adapted information to create an advertising poster for an after school exercise program.
- *Band Competition:* 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:* 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:* Students used online database tools to help them 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.

Questionnaires

After completing the two test modules, students also completed, on a computer, a 30-minute international student questionnaire. It 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 toward the use of ICT.

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

- A 30-minute *teacher questionnaire* that began by asking teachers to provide some basic background characteristics about themselves. These questions were followed by questions relating to teachers' reported use of ICT in teaching, their attitudes toward using ICT in teaching, and their participation in professional learning activities relating to their use of ICT when teaching.
- A 10-minute *ICT coordinator questionnaire* that asked ICT coordinators about the resources available in their school to support the use of ICT in teaching and learning. The questionnaire addressed both technological (e.g., infrastructure, hardware, and software) and pedagogical support (such as through professional learning).

- A 10-minute *principal questionnaire* through which principals provided information about school characteristics and then about school approaches to providing CIL-related teaching and incorporating ICT in teaching and learning.

An additional questionnaire—the national context questionnaire—was used to gather information from ICILS national researcher coordinators (NRCs) about national approaches to developing students' CIL capacity. This information included policies and practices as well as plans for using ICT in education. When answering this questionnaire, which was administered online, the NRCs also drew on the expertise of national experts to provide the required information.

Measures

The computer and information literacy construct

As stated above, computer and information literacy (CIL) refers to 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 the community" (Fraillon et al., 2013, p. 18). The CIL construct was conceptualized around two strands that framed skills and knowledge addressed by the CIL instruments.

Each strand was made up of several aspects that referenced specific content. Strand 1 of the framework, titled *collecting and managing information*, focused on the receptive and organizational elements of information processing and management. Strand 2 of the construct, titled *producing and exchanging information*, focused on using computers as productive tools for thinking, creating, and communicating. Chapter 2 of this report provides more details about the CIL construct.

The computer and information literacy scale

The Rasch item response model (Rasch, 1960) was used to derive the CIL scale from student responses to the 62 test questions and large tasks (which corresponded to a total of 81 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 properties of the task. Each large-task assessment criterion was therefore also an item in ICILS.

The final reporting 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. Plausible value methodology with full conditioning was used to derive summary student achievement statistics.

The ICILS described scale of CIL achievement 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 correctly responding to each item. Ordering the item descriptors according to their scaled difficulty (from least to most difficult) resulted in 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 difficulty.

Analysis of this item map and the student achievement data were then used to establish proficiency levels that had a width of 85 scale points and level boundaries at 407, 492, 576, and 661 scale points (rounded to the nearest whole numbers). Student scores

below 407 scale points indicate CIL proficiency below the lowest level targeted by the assessment instrument.

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 in relation to 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 information-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 information-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. They are also aware of some of the 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 communication tasks and add simple content to information products. They demonstrate familiarity with the basic layout conventions of electronic documents.

Measures based on the student questionnaire

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

- 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 questions about the following:

- Whether they had learned various ICT tasks at school;
- Their self-efficacy in using computers; and
- Their interest and enjoyment in using computers.

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 computers for general applications, school purposes, communication and information exchange, and recreation. Measures of *perceptions* included ICT self-efficacy (in relation to basic and advanced tasks) and interest and enjoyment in using ICT.

Measures based on the teacher questionnaire

A number of the measures based on the teacher questionnaire were single-item indices. Such measures included:

- Experience with using computers for teaching purposes (made up of three categories based on years of using computers in teaching); and
- Frequency of computer use at school when teaching, at school for other purposes, and outside of school (made up of five categories).

Additional sets of items were designed to generate scales reflecting the following:

- Teachers' views of ICT for teaching and learning;
- Teachers' self-confidence in using ICT; 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 *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, teachers were given the following instruction:

> 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 also 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 questionnaires

The school questionnaires 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.

Computer-based test delivery

ICILS used purpose-designed software for the computer-based student assessment and questionnaire. These were administered primarily using USB drives attached to school computers. Although the software could have been delivered via the internet, the USB delivery ensured a uniform assessment environment for students regardless of the quality of internet connections in participating schools. After administration of the student instruments, the ICILS research team either uploaded data to a server or delivered this information to national research centers for them to upload.

The teacher and school questionnaires were usually completed on computer (over the internet in order to access IEA's Data Processing and Research Center server in Hamburg, Germany). However, respondents could also complete the questionnaires on paper.

Data

Countries

Twenty-one countries participated in ICILS 2013. They were Australia, the City of Buenos Aires (Argentina), Chile, Croatia, the Czech Republic, Denmark, Germany, Hong Kong SAR, the Republic of Korea, Lithuania, the Netherlands, Norway (Grade 9), Newfoundland and Labrador (Canada), Ontario (Canada), Poland, the Russian Federation, the Slovak Republic, Slovenia, Switzerland, Thailand, and Turkey.[1] Three of these participants represented education systems within their countries—the City of Buenos Aires (Argentina), Newfoundland and Labrador (Canada), Ontario (Canada)—and were considered to be benchmarking participants. Canada originally intended to participate as a country, but ultimately only the two provinces of Newfoundland and Labrador and Ontario participated.

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. If the average age of students in Grade 8 was below 13.5 years, Grade 9 became the target population.[2]

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

1 The majority of the entities that participated in ICILS were countries. Some subunits of countries featuring a distinct education system also participated in ICILS, for example Hong Kong, a special administrative region of China. For reasons of simplicity, the text refers to both participating countries and education systems as "countries."

2 Norway decided to survey students and their teachers at the end of their ninth grade. Their results were annotated accordingly in the international report (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014).

the basis of national characteristics. However, as a guide, each country was instructed to plan for a minimum sample size of 150 schools. The sampling of schools constituted the first stage of sampling both students and teachers. The sample of schools ranged in number from between 138 and 318 across countries.

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 almost 60,000 lower-secondary students in more than 3,300 schools from 21 countries or education systems within countries. These student data were augmented by data from almost 35,000 teachers in those schools and by contextual data collected from school ICT coordinators, school principals, and national research centers.

The main ICILS survey took place in the 21 participating countries between February and December 2013. The survey was carried out in countries with a Northern Hemisphere school calendar between February and June 2013 and in those with a Southern Hemisphere school calendar between October and December 2013. 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.

Outline of the technical report

This overview of the International Computer and Information Literacy Study is followed by 12 chapters. Three chapters cover the instruments that were used in the study. Chapters 2 and 3 are concerned with the ICILS test. Chapter 2 focuses on the development of the tests, and Chapter 3 provides an account of the computer-based assessment systems. Appreciation of the material in these chapters provides an essential foundation for interpreting the results of the study. 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 NRCs.

Chapters 5 through 9 focus on the implementation of the survey in 2013. 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 achieved participation rates. 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 through 13 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 or how the responses to tasks and items were used to generate the scale scores and proficiency levels. Chapter 12 describes the analogous scaling procedures for the questionnaire items (mainly the student and teacher questionnaires). The final chapter, Chapter 13, presents an account of the analyses that underpinned the ICILS 2013 international report (Fraillon et al., 2014).

References

European Commission. (2006). *Recommendation 2006/962/EC of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning* (Official Journal L 394 of 30.12.2006). Luxembourg, Brussels: Author. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006H0962& from=EN

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.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International Computer and Information Literacy Study: Assessment framework.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

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

CHAPTER 2:

ICILS Test Development

Julian Fraillon

Introduction

The ICILS assessment was developed over a 20-month period from April 2010 to December 2012. Most of this work was conducted by the international study center (ISC) at the Australian Council for Educational Research (ACER) in collaboration with national research coordinators (NRCs) and the project advisory committee (PAC). SoNET Systems conducted the software development for the test modules.

This chapter provides a detailed description of the test development process and review procedures as well as the test design implemented for the ICILS field trial and main survey. Table 2.1 provides an overview of the test development processes and timeline.

Test scope and format

ICILS assessment framework

The ICILS student test was developed with reference to the ICILS assessment framework[1] (Fraillon, Schulz, & Ainley, 2013) and was designed to measure computer and information literacy (CIL), 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 the community" (Fraillon et al., 2013, p. 18).

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

- Strand 1, Collecting and managing information, comprising three aspects:
  - Aspect 1.1: Knowing about and understanding computer use
  - Aspect 1.2: Accessing and evaluating information
  - Aspect 1.3: Managing information.
- Strand 2, Producing and exchanging information, comprising four aspects:
  - Aspect 2.1: Transforming information
  - Aspect 2.2: Creating information
  - Aspect 2.3: Sharing information
  - Aspect 2.4: Using information safely and securely.

Although the ICILS assessment framework leaves open the possibility that CIL may comprise more than one measurement dimension, it does "not presuppose an analytic structure with more than one subscale of CIL achievement" (Fraillon et al., 2013, p. 19). In accordance with analyses of dimensionality in the ICILS student-achievement data, CIL was reported as a single scale.

1 The framework can be downloaded from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICILS_2013_Framework.pdf

Table 2.1: Test development processes and timeline

Date/Period		Activity
April to June	2010	Drafting of preliminary module ideas at ICILS international study center
June	2010	Review of proposed test-development process and preliminary module ideas and module development workshop at first meeting of national research coordinators (Amsterdam)
July to October	2010	Drafting, review, and refinement of modules at ICILS international study center
October	2010	Web-based review of test-module storyboards by national research coordinators
November to February	2010–2011	Revision of modules at ICILS international study center
February	2011	Review of proposed test modules and trial of test delivery software at second meeting of national research coordinators (Hamburg)
February to March	2011	Revision of modules at ICILS international study center
March	2011	Web-based review of test modules by national research coordinators and project advisory committee
March to June	2011	Revision of modules at ICILS international study center
June	2011	Review of modules proposed for inclusion in field trial test and confirmation of test design at third meeting of national research coordinators (Ljubljana)
July	2011	Finalization of field trial modules at ICILS international study center
July to November	2011	Preparation of field trial scoring guides for constructed-response items and large tasks
November	2011	Review of field-trial scoring guides for constructed-response items and large tasks at ICILS field trial scorer-training meeting (Hamburg)
December	2011	Revision of field trial scoring guides for constructed-response items at ICILS international study center
May	2012	Analysis of field trial item data and recommendations for modules/items to be included in main survey test (field trial analysis report) at ICILS international study center
September	2012	Review of field trial analysis report and recommendations for test design and modules/items proposed for inclusion in main survey at fourth meeting of national research coordinators (Brig)
September to November	2012	Preparation of main survey scoring guides for constructed-response items and large tasks
November	2012	Review of main survey scoring guides for constructed-response items and large tasks at ICILS main survey scorer-training meeting (Hamburg)
December	2012	Revision of main survey scoring guides for constructed-response items at ICILS international study center

The ICILS test instrument

The questions and tasks making up the ICILS test instrument were presented in four modules, each of which took 30 minutes to complete. Each student completed two modules randomly allocated from the set of four.

A module is a set of questions and tasks based on an authentic theme and following a linear narrative structure. Each module has a series of smaller discrete tasks, each of which typically takes less than a minute to complete. These tasks are followed by a large task that typically takes 15 to 20 minutes to complete. The narrative of each module positions the smaller 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, the ICILS 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 four ICILS assessment modules and large tasks.

Table 2.2: Summary of ICILS test modules and large tasks

Module	Description and Large Task
After School Exercise	Students set up an online collaborative workspace to share information and then select and adapt information to create an advertising poster for the after-school exercise program.
Band Competition	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	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	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.

The ICILS test modules included six types of task described below.

- *Multiple-choice:* Students answered these questions by clicking on a radio button on the test interface.
- *Constructed-response:* Students answered these questions by writing one or two sentences in a text field on the test interface.
- *Drag and drop:* Students answered these questions by clicking on and dragging responses on the screen into a response grid.
- *Linear skills:* Students completed these tasks by executing one or more commands in a given sequence (such as copying and pasting text or opening a hyperlink).
- *Nonlinear skills:* Students completed these tasks by executing a set of commands for which the desired outcome could be reached by using a range of command sequences (such as searching an online database by filtering for more than one feature).
- *Large tasks:* Students completed these tasks by using product-specific software to produce an information product (such as a presentation or a webpage) in an open environment. The large tasks were scored according to analytic criteria relating to the students' use of information and software functions to produce the information product.

Further details of the ICILS test modules, including example items, are presented in the ICILS assessment framework (Fraillon et al., 2013) and the ICILS international report (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014).

Test-development process

The test-development process consisted of a series of stages. Although these stages followed one another 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 the NRCs reviewed and the ISC then revised. Sometimes this process was repeated. The PAC conducted a review of the final draft materials in parallel with the final NRC review.

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

In preparation for the first meeting of the ICILS NRCs, the ISC drafted a set of three module ideas. The module ideas, drafted as concept documents, included a description of the narrative structure of each module and descriptions of the discrete and large tasks that students could complete.

Refinement of preliminary module ideas

NRCs had opportunity to discuss and review the module ideas at the first NRC meeting. They also received broader information about the test design and the software environment in which the modules would be developed. NRCs were invited to discuss each module idea and to suggest changes and also new ideas for module themes. This process took place in a module-development workshop held during the NRC meeting.

As part of the module-development workshop, participants were also introduced to the questions and issues directing the 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 bulleted 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 CIL construct 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 might 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 crossculturally 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 one 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 providing 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 appeared to be equally correct)?
  - Were the scoring criteria sufficient for scorers, and did they clearly distinguish the different levels of performance?

Development of module storyboards

After the first NRC meeting, the ISC developed storyboards for five modules. During this work, the ISC drew on the test-development expertise of other staff at ACER.

The storyboards were presented in the form of a Microsoft PowerPoint mock-up of each task/item set in sequence. The tasks were presented this way 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.

Review of module storyboards

When reviewing the draft storyboards, NRCs made recommendations relating to the modules. The ISC also asked the NRCs to recommend which four modules should be developed further. The NRCs' feedback informed further revision of the module storyboards; a module assessing general ICT skills was then removed from further development.

Preliminary enactment of draft storyboards

Once development of the module storyboards had been completed, they were sent to SoNET Systems 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 as well as the feasibility of developing them into their final form; and it enabled the functionality of the test delivery system to be tested and refined.

Face-to-face review of draft modules

One focus of the second meeting of NRCs was to review the draft test modules in their partially enacted formats. At this point in the instrument-development process, most items in most modules were fully functioning.

The draft modules were further revised on the basis of NRC feedback from this meeting. Enactment of the modules in the delivery software continued from this point on.

Online review of draft modules

Once the final draft versions of the modules had been fully enacted in the test delivery environment, they were made available for a penultimate review prior to the field trial. Screenshots of each item and task were made available to NRCs and members of the project advisory committee (PAC), who provided feedback on the modules. The modules were further revised in light of this feedback.

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

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

Field trial scoring 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.

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, refinements were recommended for a small number of tasks, and one task was removed. The NRCs also strongly recommended that all four test modules used in the field trial should be retained for use in the main survey, given that all four exhibited satisfactory validity, functioning, and measurement properties.

Post field trial revision

After the field trial, minor modifications were made to a small number of tasks and to the scoring guides. The main survey instrument, comprising four test modules, was then finalized.

Main survey scoring training

Another international scorer training meeting was conducted before the main survey. This 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

The field trial test instrument consisted of four test modules with a total of 68 tasks. 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 item types and task formats in all four modules but rather to have a mix across the modules of traditional test-format items (multiple-choice and constructed-response), skill-performance tasks, and large tasks that required students to use multiple software applications to generate a digital product. Table 2.3 shows the composition of the field trial test modules by task and item type. Overall, 27 percent of the items were based on traditional types of test items, 29 percent involved skill performance, and 44 percent were based on large tasks.

Whether tasks were scored by the computer delivery system or by trained human scorers depended on the type of task. Most tasks corresponded to a single item (with one or more score points), but each criterion for the large tasks was also an item. Table 2.4 shows the composition of the field trial test instrument by task type and associated item score points. In total, the 68 field trial tasks yielded 87 score points for inclusion in the item analysis and scaling.

The test modules were delivered to students in a fully balanced complete rotation. Table 2.5 shows this design. Altogether, there were 12 different possible combinations of module pairs. Each module appeared in six of the combinations—three times as the first and three times as the second module when paired with each of the other three. The module combinations were randomly allocated to students. Each test completed by a student consisted of two of the four modules.

This test design made it possible to assess a larger amount of content than could be completed by any individual student and was necessary to ensure a broad coverage

Table 2.3: Field trial test instrument module composition by task

Module	Task Format						
	Multiple-choice	Constructed-response	Drag and drop	Linear skills	Nonlinear skills	Large task criterion	Total
After School Exercise	3	4	0	3	1	9	**20**
Band Competition	1	2	2	2	3	7	**17**
Breathing	0	6	0	3	0	8	**17**
School Trip	2	0	0	5	1	6	**14**
Total	**6**	**12**	**2**	**13**	**5**	**30**	**68**

Table 2.4: Composition of the field trial test instrument by task type and score points

Task Type/Item Format	Item Scoring (Computer/Human)			Score Points	
	Computer scored	Human scored	Total	Total score points	Percentage of score points
Multiple-choice	6	0	6	6	7
Constructed-response	0	12	12	13	15
Drag and drop	2	0	2	3	3
Linear skills	13	0	13	13	15
Nonlinear skills	5	0	5	5	6
Large task criterion	1	29	30	47	54
Total	**27**	**41**	**68**	**87**	**100**

of the content of the ICILS assessment framework. This design also controlled for the influence of item position on difficulty across the sampled students and provided a variety of contexts for the assessment of CIL. The term *booklet* in Table 2.5 refers to each combination of modules that was used in the field trial.

Table 2.5: Field trial module combinations

Combination	Position	
	1	2
1	A	H
2	A	B
3	A	S
4	H	A
5	H	B
6	H	S
7	B	A
8	B	H
9	B	S
10	S	A
11	S	H
12	S	B

Note:
A: After School Exercise
B: Band Competition
H: Breathing
S: School Trip

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 item mapping to assessment framework

Framework Aspect		Number of Items/Tasks	Percentage of Task	Number of Score Points	Percentage of Score Points
1.1	Knowing about and understanding computer use	10	15	10	11
1.2	Accessing and evaluating information	13	19	16	18
1.3	Managing information	5	7	5	6
Total	**Strand 1**	**28**	**41**	**31**	**36**
2.1	Transforming information	14	21	18	21
2.2	Creating information	15	22	25	29
2.3	Sharing information	1	1	1	1
2.4	Using information safely and securely	10	15	12	14
Total	**Strand 2**	**40**	**59**	**56**	**64**
Total	**All aspects**	**68**	**100**	**87**	**100**

As stated in the ICILS assessment framework, "… the 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., 2013, p. 43). The intention that the two strands would be adequately represented in the test was achieved, although approximately twice as many score points related to Strand 2 as to Strand 1. These proportions corresponded to the amount of time the ICILS students were expected to spend on each strand's complement of tasks. The first three aspects of Strand 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.

Selection of items for the 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 12 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 the removal of one task. 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 four field trial test modules for use in the main survey.

Main survey test design and content

Test design

The main survey test instrument consisted of four test modules with a total of 70 tasks. Some of these tasks generated a number of score points based on the criteria that were applied to the large tasks. Table 2.7 shows the composition of the main survey test modules by task and item type.

Table 2.8 shows the composition of the main survey test instrument by task type and associated item score points. The 70 main survey tasks yielded 98 score points for inclusion in the item analysis and scaling. Overall, 17 percent of the score points were based on traditional types of test items, 23 percent involved skill performance, and 59 percent were based on large tasks. Please note that the numbers reported in Tables 2.7 and 2.8 show the tasks, items, and score points *before* completion of the main survey and analysis. The reduced numbers shown in the ICILS international report (Fraillon et al., 2014) are those that were evident after the item selection and scaling analysis. Details of these analyses are included in Chapter 11.

Table 2.7: Main survey test instrument module composition by task

Module	Task Format						
	Multiple-choice	Constructed-response	Drag and drop	Linear skills	Nonlinear skills	Large task criterion	Total
After School Exercise	3	4	0	3	1	11	**22**
Band Competition	1	2	2	2	3	7	**17**
Breathing	0	3	0	3	0	10	**16**
School Trip	2	0	0	5	1	7	**15**
Total	**6**	**9**	**2**	**13**	**5**	**35**	**70**

Table 2.8: Composition of the main survey test instrument by task type and score points

Task Type/Item Format	Item Scoring (Computer/Human)			Score Points	
	Computer scored	Human scored	Total	Total score points	Percentage of score points
Multiple-choice	6	0	6	6	6
Constructed-response	0	9	9	11	11
Drag and drop	2	0	2	3	3
Linear skills	13	0	13	13	13
Nonlinear skills	5	0	5	7	7
Large task criterion	1	34	35	58	59
Total	**27**	**43**	**70**	**98**	**100**

A comparison of Tables 2.4 and 2.8 shows that the larger number of score points available for analysis in the main survey in comparison to the field trial was mainly a result of the additional criteria and score points attributable to the large task criteria. These additional criteria and score points for the main survey were a product of the analyses of the field trial data and consultations with NRCs regarding the scoring of the large tasks in the field trial.

Students received the test modules in the same fully balanced complete rotation that was used in the field trial. Table 2.9 shows this design for the main survey. As before, the term *booklet* refers to each combination of modules used in the main survey.

Table 2.9: Main survey module combinations

Combination	Position	
	1	2
1	A	H
2	A	B
3	A	S
4	H	A
5	H	B
6	H	S
7	B	A
8	B	H
9	B	S
10	S	A
11	S	H
12	S	B

Note:
A: After School Exercise
B: Band Competition
H: Breathing
S: School Trip

Main survey coverage of the CIL framework

All main survey items were developed according to and mapped against the ICILS CIL framework. Table 2.10 shows this mapping.

A comparison of Tables 2.6 and 2.10 reveals that the final test instrument provided very similar CIL framework coverage to that of the field trial instrument. The slightly higher proportion of items in Strand 2 than in Strand 1 in the main survey instrument compared to the field trial instrument was a result of the increase in the number of large-task criteria and associated score points from the field trial to the main survey. As in the field trial, roughly twice as many score points related to Strand 2 as to Strand 1.

Table 2.10: Main survey item mapping to assessment framework

Framework Aspect		Number of Items/Tasks	Percentages of Tasks	Number of Score Points	Percentage of Score Points
1.1	Knowing about and understanding computer use	10	14	10	10
1.2	Accessing and evaluating information	10	14	14	14
1.3	Managing information	5	7	7	7
Total	**Strand 1**	**25**	**36**	**31**	**32**
2.1	Transforming information	16	23	23	23
2.2	Creating information	18	26	31	32
2.3	Sharing information	1	1	1	1
2.4	Using information safely and securely	10	14	12	12
Total	**Strand 2**	**45**	**64**	**67**	**68**
Total	**All aspects**	**70**	**100**	**98**	**100**

Released test module

One test module, *After School Exercise*, has been released since publication of the ICILS international report (Fraillon et al., 2014). This module required students to set up an online collaborative workspace to share information and then to select and adapt information so as to create an advertising poster for the after school exercise program.

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 86 to 94 of the ICILS international report, while a demonstration video of the module can be found on IEA's website: http://www.iea.nl/index.php?id=475

Summary

The test development process for ICILS was guided by the ICILS assessment framework and was carried out in an iterative process of development, refinement, and review over a period of 20 months beginning in early 2010. Test development staff at the ISC were responsible for drafting and refining the test modules. Storyboards of the modules and tasks were first developed as static presentations in Microsoft PowerPoint and then, after review and refinement, were authored as fully functioning tasks in the ICILS test delivery software environment. National center staff in participating countries provided detailed review feedback at a number of key points during test development, and this information informed the refinement of the test modules and tasks throughout the process.

The four ICILS test modules were designed to assess the breadth of content described in the assessment framework. The intention to have the two strands adequately represented in the test was achieved, although approximately twice as many score points related to Strand 2 as to Strand 1. These proportions corresponded to the amount of time the ICILS students were expected to spend on each strand's complement of tasks. The first three aspects of Strand 2 were assessed primarily via the large tasks at the end of each module; students were expected to spend roughly two thirds of their working time on these tasks.

Each test completed by a student consisted of two of the four modules. This test design made it possible to assess a larger amount of content than could be completed by any individual student and was necessary to ensure a broad coverage of the content of the ICILS assessment framework. The modules were presented across students in a fully balanced complete rotation to account for any potential order and module combination effects influencing the student achievement data. One of the four test modules, After School Exercise, was released after publication of the ICILS international report. Details of the module are included in the ICILS international report, and a video demonstration of the module is available on the IEA website.

References

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.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International Computer and Information Literacy Study assessment framework.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

CHAPTER 3:

Computer-Based Assessment Systems

Julian Fraillon and Ralph Carstens

Introduction

ICILS was the first IEA study to collect 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. 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.

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 elements 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 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 at the IEA Data Processing and Research Center (DPC), 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 SurveySystem (IEA OSS), which supports the translation, adaptation, and subsequent delivery of computer-based questionnaire material; and
- 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.

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

The key technologies used in these software tools are the Microsoft.NET framework (including ASP.NET), Microsoft SQL Server (full or compact), Microsoft Access databases, and Microsoft IIS (Internet Information Services).

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

Components developed for ICILS

The ICILS international study center (ISC) at the Australian Council for Educational Research (ACER) contracted SoNET Systems to adapt (and in some cases develop) the software components directly relating to the collection of student data via computer. These software components are part of a broader suite known as AssessmentMaster. The software modules specifically relating to the delivery of the ICILS student test and questionnaire include an administration module, a translation module, a delivery engine module, and a scoring module.

The key technologies used in these software components were Apache webserver, PHP, MySQL, and Firefox (portable or desktop version).

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. The IEA DPC, in close cooperation with the entire international team, has developed and refined instructions and guidelines for IEA WinW3S, IEA OSS, and IEA DME 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, 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 the administrators who were administering the tests in schools. All user manuals were developed for use in the field trial and refined after it for use in the ICILS main survey.

ICILS system architecture

Table 3.1 provides an overview of the ICILS 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 target populations. As can be seen, ICILS 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

Developing the delivery engine

Early in the planning stages of ICILS, the international research team decided to use an existing test delivery system rather than develop a new system from the ground up. An existing system at this time was AssessmentMaster, earlier versions of which had been used to deliver national sample assessments in Australia. However, when work on ICILS began in early 2010, AssessmentMaster offered only a delivery engine and scoring module, which meant a translation module and administration module had to be developed for use in ICILS. Some adaptations to the delivery engine and scoring module were also necessary to ensure they suited the study's requirements.

Development of the computer-based delivery platform took place in parallel with test development (see Chapter 2 for details of this development). Once the test modules storyboards had been completed, they were sent to SoNET Systems for authoring

Table 3.1: Tabular overview of ICILS computer-based or computer-supported system 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 paper-based questionnaires	Microsoft Word, direct authoring of paper-based questionnaires
Translation	AssessmentMaster translation module, web-based	IEA OSS translation module (from paper materials): desktop-based	Microsoft Word: desktop-based
Sample selection and ID provisioning	IEA Within-School Sampling Software (IEA WinW3S)		
Delivery systems	AssessmentMaster delivery module	IEA OSS delivery module: web-based	Personalized questionnaire prints
Initialization	Student ID, password, and language information from WinW3S	Respondent ID and password from WinW3S	Labels with respondent ID from 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 capture using IEA DME software after administration (all respondents)
Data merging (across respondents)	Upload to AssessmentMaster admin module	n/a	Merging from multiple DME databases (where used)
Data scoring	AssessmentMaster scoring module	n/a	n/a
Data management	IEA WinW3S; crosscheck of expected versus available data		

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 development process for the field trial took place between October 2010 and June 2011, although preparatory work (such as decisions about which tasks and functionality would be viable) began in early 2010. One critical task during this test design process was determining whether each task and software feature was sufficiently viable to undergo translation. Although the translation module was developed in parallel with development of the test content, translation viability was deemed to be an important part of the test development process.

The test delivery engine was web based, which meant that technically the tests could be delivered over the internet. However, the ICILS 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 materials were developed, rendered, and tested using only Mozilla Firefox. For this reason, the research team recommended that the translation, scoring, and

administration modules be accessed only through Firefox. 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 was practical after each test session.

An alternative delivery option was made available after the field trial. Under this system, the delivery engine was installed on a notebook computer connected to a school local area network (LAN). The school could then access the test locally over the school LAN via Mozilla Firefox.

School coordinators were provided with the following list of minimum specifications for the ICILS tests.

- A minimum of 1GB RAM;
- At least one functioning USB 2.0 port (or faster);
- A monitor and video card that could support a screen resolution of 1024x768 pixels (note that audio was not used in the ICILS assessment);
- Any of the following operating systems: Windows 2000/XP/Vista/7/Win Server 2003/ Win Server 2008 *or* Macintosh OS or Linux with reputable Windows emulation software (required to ensure the student assessment could be administered using USB sticks) *or* Mac OS/Linux, which could be connected to a (e.g., the school) LAN and had either Firefox 6.0 or higher installed or the ability to install Mozilla Firefox 6.0 or higher prior to the day of testing (required for the student assessment to be administered using a single laptop web-server connected to the LAN); and
- A minimum screen size of [15 inches].

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 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; and
- A countdown timer for students per module.

Developing the translation module

The translation module was developed through consultation among colleagues at the ISC at ACER, the IEA DPC, the IEA Secretariat, the company contracted to complete translation verification (cApStAn), and the national research coordinators (NRCs). The first prototype of the translation module, developed in early 2011, underwent an internal review. It was further reviewed at a mid-2011 NRC meeting as part of its preparation for use in the field trial. Several refinements were made to the functionality of the translation module after the field trial. These included the addition of new features and revisions to the responsibilities and privileges of each of the user roles.

Because the translation module was a web-based application 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;
- Mozilla Firefox web-browser Version 3.6 or higher (the translation platform was developed to work with Mozilla Firefox and tested using this web-browser).

Development of the translation module 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;
- Opportunity for translators, reviewers, and verifiers to monitor the progress of task completion (i.e., translation state).

Developing the scoring module

The ICILS scoring module was an adaptation of an existing scoring module. This adaptation included the development of some new features for use in ICILS. 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 did not require scorers or scoring trainers to be proficient in English.)

The functionality of the scoring module, a web-based application that could be accessed through a web-browser, was refined following the field trial.

In order to access the application, users needed:

- A computer with a broadband or equivalent high-speed internet connection;
- Mozilla Firefox web-browser Version 3.6 or higher.

The scoring module 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, 20%) of student responses to be blind double-scored for the purpose of monitoring inter-rater reliability;
- 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;
- Enter a "training" mode in order to score pre-scored work and to receive feedback on scoring accuracy.

In addition, team leaders could:

- Review (check-score and amend) scorers' scores;
- 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;
- View interscorer reliability reports.

Developing the administration module

The ICILS administration module 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 WinW3S.

As occurred with the scorer module, the functionality of the administration module was refined after the field trial. It, too, 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;
- Mozilla Firefox web-browser Version 6 or higher.

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

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

Challenges with computer-based delivery in ICILS

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 are 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 computer-based assessments 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 test-delivery platform, including its USB-delivery option, was designed to minimize the potential for variation in students' test-taking experience. The onscreen appearance of the ICILS 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 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, the inspection of unweighted yet cleaned data from all 21 participating education systems (67,771 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, and 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 provided across all countries given the per-country proportions were similar.

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

Participation Status	Test	Questionnaire
Left school permanently	1.1%	
Parental permission denied	2.2%	2.5%
Absent	6.4%	6.6%
Incompatible or failed equipment before assessment	0.2%	
Technical failure during assessment	0.4%	0.5%
USB stick lost or upload failed after assessment	0.8%	
Participated	88.9%	88.3%
Total	100%	

Table 3.2 shows that test and questionnaire data were collected from 88.9 percent and 88.3 percent of sampled students respectively. The slightly lower percentage of data collected for the questionnaire is a result of three factors:

1. All students answered the questionnaire after they had completed the test, so there was marginally more chance for the testing system to fail (0.1% more technical failures occurred during administration of the questionnaire than administration of the test).
2. 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 (0.2% more students were absent for the questionnaire than for the test).
3. In some countries parents gave permission for their children to complete the test but not the questionnaire (resulting in a further absentee difference of 0.3%).

Roughly 7.5 percent of sampled students overall had either permanently left school (1.1%) 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 for the students concerned. These cases 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. A small number of NRCs mentioned slow or sometimes even frozen test systems as a result of old computers or interacting applications (e.g., Skype). However, none of the NRCs indicated or reported a lack of suitable delivery options (primary plus two alternatives) 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) accounted for a level of nonresponse many times larger than that due to technical issues.

Of the technical problems, cases involving lost USB sticks, corrupted data, or data that could not be uploaded constituted the largest factor (0.9% of all sampled students). However, these cases were not distributed uniformly: 6 of the 21 education systems contributed more than three-quarters of cases here. We need to note that national centers were responsible for selecting and purchasing the USB drives. The ISC instructed national centers to purchase USB drives produced by reputable brands and to test the system across a range of sticks, but it did not require them to inform the international center of their choice or submit their USB drives for testing.

Summary

ICILS was the first IEA assessment in which the student instruments were delivered entirely on computer. The ICILS team used a pre-existing system for authoring and delivering the student instruments and integrated it with existing IEA-developed software systems developed by IEA and used in other studies. This integration proved to be highly successful. The overall data loss or corruption was minimal, especially in contrast to the much larger factors relating to school and/or parental cooperation or student absences. The ICILS team therefore concluded that these factors had the greater impact on survey response and arose out of general factors or the survey topic/importance rather than technology and/or delivery mode.

CHAPTER 4:

ICILS Questionnaire Development

Wolfram Schulz and John Ainley

Introduction

In this chapter, we describe the development of the international questionnaires for students, teachers, school principals, information and communication technology (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. The teacher questionnaire was designed to gather teachers' perspectives on the general school environment for ICT use, their familiarity with using ICT, their attitudes regarding the use of ICT in teaching, and their use of ICT for teaching and learning in a reference class.

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 (such as through professional learning).

An online questionnaire—the national contexts survey—for the national research coordinators (NRCs) was designed to collect contextual information at the national (or subregional) 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, Schulz, & Ainley, 2013). The assessment framework consisted of two parts:

- *The computer and information literacy framework:* This outlined the outcome measures addressed through the cognitive test and those parts of the student questionnaire designed to measure student perceptions.
- *The contextual framework:* This mapped the context factors expected to influence outcomes and to explain their variation.

The contextual framework identified the context variables that reflect the environment in which learning computer and information literacy (CIL) takes place. It assumes that young people develop their understandings of ICT through a number of activities and experiences that take place not only in the school and the classroom but also in the home and other places outside school.

Students' development of CIL is influenced by what happens in their schools and classrooms (the instruction they receive, ICT availability and use within the school context, ICT use for teaching and learning), their home environments (socioeconomic background, availability and use of ICT at home), and their individual characteristics. The latter shape the way students *respond* to learning about computers and computing.

Contextual influences on CIL learning are conceived as either antecedents or processes. Antecedents refer to the general background that affects how CIL 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 learning more directly (e.g., the extent of opportunities for CIL learning during class, teacher attitudes toward ICT for study tasks, and students' computer use at home). The basic classification and organization of antecedent and process-related contextual factors that influence CIL outcomes are illustrated in Appendix B of this report.

Reference to a general conceptual framework made it possible to locate variables collected through contextual instruments in a two-by-four grid, where antecedents and processes constituted columns and the four levels constituted the rows. Table 4.1 shows this grid with examples of measures in the appropriate cells. 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. The national contexts survey and published sources provided variables at the system or national level.

Table 4.1: Mapping of ICILS context variables to framework grid

Level of ...	Antecedents	Processes
Wider community	**NCS and other sources:** Structure of education Accessibility of ICT	**NCS and other sources:** Role of ICT in curriculum
School/classroom	**ScQ and TQ:** School characteristics ICT resources	**ScQ and TQ:** ICT use in teaching
Student	**StQ:** Gender Age	**StQ:** ICT activities Use of ICT
Home environment	**StQ:** Parental SES ICT resources	**StQ:** Learning about ICT at home

Key: NCS = national contexts survey; StQ = student questionnaire; ScQ = school questionnaire; TQ = teacher questionnaire.

Development of the ICILS context questionnaires

The international study center (ISC) at the Australian Council for Educational Research (ACER) coordinated, in liaison with their partner institutions, the development and implementation of the ICILS context questionnaires for students, teachers, and schools. Several national centers also proposed additional item material for the questionnaires. The development work included extensive reviews and discussions at different stages of the process with experts from the ICILS project advisory committee (PAC) and national centers.

The development process for the student, teacher, and school questionnaires followed a sequence which paralleled that for the ICILS test outlined in Table 2.1. Specifically, the questionnaire development involved three phases:

- Phase 1: This phase, from January 2010 to June 2011, included the development of field trial material guided by the assessment framework. It also included various rounds of consultations and reviews by NRCs and PAC members (as shown in Table 2.1).
- Phase 2: This phase, from July 2011 to June 2012, involved the international field trial, conducted in 20 participating ICILS countries,[1] and subsequent analyses of field trial data to inform judgements about the suitability of questionnaire material for the main survey. The sequence of activities was the same as that shown in Table 2.1.
- Phase 3: During this final phase, which took place from July to October 2012, ISC staff discussed the field trial results with staff in the national centers and with PAC members. Phase 3 concluded with a final selection of main survey items.

During each of these phases, the procedures and criteria used to review the field trial material and results were the same for all context 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 selection of item material for the main survey:

- Relevance with regard to the ICILS 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 to develop the other context questionnaires.

Development of the student questionnaire

Students were asked to answer the ICILS student questionnaire via a computer after they had completed the CIL test. The student questionnaire had two parts: the first collected information about the students' characteristics and their home background; the second focused on the students' use of and familiarity with ICT. ICILS researchers at ACER coordinated the development and implementation of the student questionnaire

1 Two of these countries, Israel and Spain, administered the field trial but did not participate in the ICILS main survey.

in liaison with their partner institutions. Throughout the development process, work on this instrument included extensive reviews and discussions with PAC experts and staff of the national research centers.

The ICILS field trial student questionnaire material included a total of 26 questions with 132 items and was administered to samples consisting of 8,895 students from 503 schools in 20 participating countries. On average in each country, about 440 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 crossnational validity of the measures derived from the ICILS questionnaires (see examples from ICCS 2009 in Schulz, 2009). Staff then discussed the field trial outcomes and the proposed draft student questionnaire for the final data collection with NRCs and PAC experts. Their feedback helped the ISC staff further develop the questionnaire for use during ICILS' final data-collection stage.

The final international student questionnaire consisted of 26 questions containing 102 items, to be completed within the targeted time of 20 to 25 minutes. Twenty of the items were designed to capture student-background information; 82 were designed to measure students' use of and familiarity with ICT. The main survey student questionnaire consisted of the following four 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 and their parents.
- *Your use of computers and internet:* These questions asked students to report on their experience with ICT, their use at different locations, and their use of ICT applications for different purposes at home and at school.
- *Your thoughts about using computers:* These questions were designed to measure students' beliefs about their ability to do computer tasks (ICT self-efficacy) and their attitudes toward computers and computing.

The items in the student questionnaire items were designed in a way that allowed them to be administered on computer. The computer delivery platform enabled a better control of student responses (e.g., not allowing invalid responses to certain questions) and more focused filtering. For example, 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 or previous occupations.

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, and teacher views on and confidence in using computers. ISC staff at ACER coordinated development and implementation of the questionnaire and asked experts from the PAC and 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 (and that was the average time taken).

The questionnaire included a question about whether teachers used ICT in their teaching and learning. Those teachers who said yes were asked to name one class as their "reference class" and to provide information about the extent of ICT use in this class for different purposes and activities.[2]

The field trial teacher questionnaire consisted of 18 questions with a total of 128 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 5,953 teachers from 505 schools in 20 participating countries. On average, the field trial teacher samples consisted of about 300 teachers in each participating country.

The final teacher questionnaire (the one used in the main study) consisted of 16 questions with 121 items and was divided into the following five sections:

- *About you:* These questions concerned teachers' background characteristics.
- *Your use of ICT:* These questions focused on teachers' use of ICT and their confidence in doing 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 answer four further questions. These focused on use of ICT applications, use of ICT for both activities and practices, and the emphasis given to the development of ICT-based capabilities.
- *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 ICT in their respective schools.
- *Learning to use ICT in teaching:* This section asked teachers about ICT-related professional development activities and cooperation in their schools.

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's ICT coordinator. Both questionnaires were delivered online by default, but an alternative paper-based version was available in cases where respondents were unable 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.

2 For reasons of randomization and thus to ensure that the selection was not biased, teachers were asked to define the reference class as the first regular class at the target grade they had taught on or after the last Tuesday before answering the questionnaire.

The questionnaire for school principals was designed to be completed in 15 to 20 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 shorter and could be answered in 10 to 15 minutes. It included predominantly factual 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 20 questions with a total of 194 items and was administered to the principals of 505 schools from the 20 countries that participated in the field trial; 473 completed questionnaires were received. In most countries, about 24 school principals provided responses to the field trial questionnaire. The ICT coordinator questionnaire consisted of 13 questions with a total of 65 items and was completed by 480 ICT coordinators at participating schools in 20 countries.

The analyses of field trial data focused on providing empirical evidence that would assist selection of the main survey material. The relatively small number of responses in each of the participating countries (the maximum could be 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 and PAC experts before selecting the items that would be included in the main (final) survey instrument. Revisions made after the field trial included rewording some items.

The review of the field trial outcomes led to a considerable reduction in the size of the school questionnaire. The final form consisted of 16 questions with a total of 80 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 13 questions with a total of 62 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 are strongly influenced by factors 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 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 conducted as part of the Second Information Technology in Education Study (SITES) 2006 (Plomp, Anderson, Law, & Quale, 2009) and by the U.S. Department of Education (2011) informed the development of the national contexts survey.

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

The development and implementation work consisted of four phases:

- *Phase 1:* During this first phase, which spanned June to December 2010, the ISC team, in discussion with the national centers and PAC experts, reached agreement on the nature and scope of the survey's contexts and questions.
- *Phase 2:* During this phase, which encompassed January 2011 to May 2013, ISC staff, national center staff, and PAC members discussed the various draft versions of the survey and eventually reached agreement on a final version.
- *Phase 3:* Between July 2013 and January 2014, the NRCs answered the national contexts survey.
- *Phase 4:* The final phase took place between January 2014 and May 2014. During it, 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 with regard to the ICILS assessment framework;
- Relevance and additional value of gathering information about the wider community context of CIL education;
- Appropriateness for the national contexts of the participating countries;
- Validity in terms of comparability, analysis, and reporting.

The final version of the national contexts survey was placed, along with accompanying notes for guidance, online via servers at the IEA Data Processing and Research Center (DPC) in Hamburg, Germany. 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 with 106 items. The questions asked respondents about key antecedents and processes in relation to CIL 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 (ISCED 2);

- 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.

Summary

The ICILS assessment framework (Fraillon et al., 2013) was the principal basis for development of the ICILS student, teacher, school, and national center questionnaires because it identified the content to be measured in each survey.

The student, teacher, and school surveys were developed in a multistage process that included an international field trial in 20 countries as well as extensive discussions with the ICILS national research centers and experts from the ICILS PAC.

It is important for those developing crossnational survey instruments to maximize input from the wide range of stakeholders. Contributions from national center staff were crucial to developing reliable and valid ICILS instruments. NRCs reviewed proposed questionnaire material during several rounds of written consultations as well as in plenary discussions at face to face meetings.

Data on the national contexts of the countries participating in ICILS were collected via an online questionnaire. The ISC asked the staff in the national centers responsible for completing the questionnaire to draw on available national expertise in their countries during this work. The national contexts survey provided a rich database of country-level information on the context for CIL education in each country.

References

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International Computer and Information Literacy Study: Assessment framework*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

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.

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

CHAPTER 5:

Translation and Verification of ICILS 2013 Instruments

David Ebbs and Tim Friedman

Introduction

Staff at the ICILS international study center (ISC), in close collaboration with staff in the participating countries' ICILS national centers, developed an international English version of the ICILS assessment and questionnaires. Countries subsequently translated and adapted these materials to their languages of instruction. Throughout this process, the overarching aim was to create high-quality instruments that were internationally comparable yet also appropriate to each country's national context and education system. The detailed guidelines on translation and adaptation provided to all participating countries are described in the *ICILS 2013 Survey Operations Procedures, Unit 3* (ICILS International Study Center, 2012).

The ICILS instruments were administered in 22 languages. English-language instruments were administered in three countries, and Spanish, French, and German versions were administered in two countries each. Even in those countries where the language of testing was English, adaptations were still required to suit the cultural setting and the version of English used.

Given that high-quality translations and adaptations were crucial to the crossnational comparability and validity of the ICILS data, all national instruments underwent a stringent international verification process. This process was overseen by the ISC at the Australian Council for Educational Research (ACER), the IEA Secretariat, and the IEA Data Processing and Research Center (DPC). The particular aim of this quality-assurance process, which included a thorough review of adaptations, translation, and layout, was to ensure that national versions were equivalent across countries to the greatest extent possible.

The ISC managed the review of adaptations to all instruments and verification of the layout of the student instruments and the paper-based versions of the teacher, ICT coordinator, and principal questionnaires. The IEA Secretariat coordinated the translation verification of all instruments, and the IEA DPC managed the layout verification of the online teacher, ICT coordinator, and principal questionnaires. Participating countries were requested to submit materials for verification before both the field trial and the main survey data collections.

In general, countries complied very well with the verification requirements. All participants submitted their instruments for the three phases of instrument verification for the field trial and for the main survey.

Translation of ICILS 2013 instruments

ICILS 2013 instruments requiring translation and adaptation

The ICILS 2013 instruments requiring translation and/or adaptation were:

- The student cognitive test (online delivery);
- The student questionnaire (online delivery); and
- The questionnaires (including instructions and covers) for teachers, ICT coordinators, and school principals (paper-based delivery and optional online delivery).

The ICILS 2013 manuals and guides were also translated where necessary. These resources included the following:

- The school coordinator and test administrator manuals; and
- The scoring guides for constructed-response items and large tasks.

Of these, the survey instruments (cognitive test and questionnaires) were subject to the international verification procedure. The ISC provided participating countries with electronic files of all materials to be adapted and/or translated. In addition, the cognitive test and questionnaire items were listed in a single combined document—the national adaptations form, which the national research coordinators (NRCs) used to register their adaptations to the instruments. Reviewers of the survey instruments listed suggestions for changes and then asked the NRCs to respond to them.

Languages used in ICILS 2013

For most participating countries, identifying the language that would be used for testing (the target language) was straightforward, as it was typically the dominant language used in public and private arenas of society, including the education system. However, in some countries, there was more than one official language or language of instruction in schools. In these cases, countries prepared instruments in all required languages. Five countries administered all or parts of the assessment (most commonly, the student instruments) in two or more languages. Table 5.1 shows the list of languages used for the ICILS survey.

Participating countries were strongly encouraged to hire qualified and experienced translators and reviewers to work with the ICILS materials. National centers were expected to enlist at least one translator (preferably certified) per target language who had the following qualifications:

- Excellent knowledge of English;
- Target language as a native language;
- A good level of computer literacy, including familiarity with technical ICT terminology relating to computers, software, and internet-based information resources and communications environments;
- Experience working with students in the target grade; and
- Familiarity with test development.

Table 5.1: Languages used for the ICILS 2013 survey instruments

Country	Language	Instruments				
		Student test	Student questionnaire	Teacher questionnaire	School questionnaire	ICT coordinator questionnaire
Australia	English	•	•	•	•	•
City of Buenos Aires, Argentina	Spanish	•	•	•	•	•
Canada (Newfoundland and Labrador, Ontario)	English	•	•	•	•	•
	French	•	•	•	•	•
Chile	Spanish	•	•	•	•	•
Croatia	Croatian	•	•	•	•	•
Czech Republic	Czech	•	•	•	•	•
Denmark	Danish	•	•	•	•	•
Germany	German	•	•	•	•	•
Hong Kong SAR	Chinese (Simplified)	•	•	•	•	•
	Chinese (Traditional)	•	•	•	•	•
	English	•	•	•	•	•
Korea	Korean	•	•	•	•	•
Lithuania	Lithuanian	•	•	•	•	•
Netherlands	Dutch	•	•	•	•	•
Norway	Bokmål	•	•	•	•	•
	Nynorsk	•	•	•	•	•
Poland	Polish	•	•	•	•	•
Russian Federation	Russian	•	•	•	•	•
Slovak Republic	Slovak	•	•	•	•	•
	Hungarian	•	•			
Slovenia	Slovene	•	•	•	•	•
Switzerland	French	•	•	•	•	•
	German	•	•	•	•	•
	Italian	•	•	•	•	•
Thailand	Thai	•	•	•	•	•
Turkey	Turkish	•	•	•	•	•

Reviewers were given the task of assessing the readability of the translation for the target population. Reviewers were required to have the following qualifications:

- Excellent knowledge of English;
- Target language as a native language;
- Knowledge of and experience in the country's present cultural context; and
- Experience working with students in the target grade.

Countries that administered the assessment in more than one target language were advised to employ a professional competent in all the target languages and therefore able to ensure that adaptations were implemented consistently in the different language versions. National centers could hire more than one translator/reviewer per language (for instance, one person to translate the test, another person to translate the questionnaires), but were responsible for maintaining the consistency of the translations and adaptations within and across instruments.

Guidelines for translation and adaptation of the instruments

The guidelines for translation and adaptation provided to all countries were designed to ensure the international comparability of the national versions of the instruments, while allowing for cultural adaptations when necessary. The guidelines were as follows (ICILS International Study Center, 2012, pp. 24–25):

- The translated text should have the same register (language level and degree of formality) as the source text.
- The translated text should use language and vocabulary appropriate for Grade 8 students.
- The translated text should have correct grammar and usage (e.g., subject/verb agreement, prepositions, verb tenses, etc.).
- The translated text should not clarify or take out text from the source text 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 and technological (e.g., computer/ICT/internet) terminology appropriate for the target language and used at this level of education.
- Idiomatic expressions should be translated appropriately, not necessarily in a literal way.
- Idiomatic names of websites or software (such as 'WebDocs') and computer-related terms (such as 'hyperlink' or 'drag and drop') should be adapted where appropriate. It is expected that in some cases no adaptation will be necessary.
- Spelling, punctuation and capitalization in the target text should be appropriate for the target language and the country's cultural context.

All of the instruments required some kind of adaptation and were therefore subject to a careful documentation and review procedure. The overarching principle of the translation and adaptation process was that the meaning and difficulty of the questions, instructions, and tasks in the instruments should be equivalent across all countries after completion of the adaptation and translation work.

Adaptation of the instruments

A number of modifications beyond those necessitated by translation into the target language were required and allowed in the specific area of computer and information literacy (CIL). They included adaptations of terms such as the names of organizations or institutions that, rather than being directly translated, needed to be replaced by the equivalent term for the local national context. The goal of such adaptations was to make the questions equally familiar to all respondents, while maintaining the same meaning and level of difficulty.

It was important that the cognitive items not be simplified, clarified, or adapted in such a way as to provide students with a hint or definition of a term that was not given in the international English version and could therefore help them identify a correct answer. For example, if a task required students to identify why a hyperlink to a web address was not safe to click on, it was essential that the translation did not provide them with any information other than that contained in the source instrument. It was also important for adaptations to be implemented consistently within instruments and, in particular, that the correspondence of text in an item stem and in its multiple-choice response options were maintained.

The international version of the materials included indications of where adaptations were required; any words in square brackets needed to be replaced with the country-appropriate term. NRCs were instructed to adapt certain recurring base expressions

from the questionnaires according to the particular country context. For example, [country of test] was to be replaced with the name of the participating country, and [target grade] was to be replaced with the name of the specific target grade in that country. Generic ISCED levels in the student questionnaire needed to be adapted to the equivalent educational terms for each country.

Some references to names of people and fictional places/countries (e.g., [Male Name A], [M-town]) were also specifically designated for adaptation. These references were adapted to target-language names that were similar in length, familiarity, and complexity to the names in the international version, the aim being to convey the same meaning and style of text as in the international version. When adapting fictional names of countries or towns, translators were explicitly instructed not to use the names of real places or countries so that students' responses would not be influenced by their knowledge or perceptions of them.

Fictional names of software and technologies also required adaptation in some instances. For example, the test contains a reference to a collaborative web-based document editor called [WebDocs]. National centers needed to determine whether this name would make sense to students or whether it required adaptation to an equivalent name (i.e., an adaptation that brought together "Web-based" and "Documents").

In principle, words not written in brackets were not to be adapted unless they were deemed inappropriate for the national context. Modifications could be made when necessary to adapt national conventions, such as measurement units and punctuation.

NRCs were provided with detailed notes on all required adaptations. These notes clarified what the particular questions were asking so that translators could select the appropriate word or expression to convey the intended meaning.

Participating countries were permitted to add a limited number of national items or categories to the questionnaires. No national additions were allowed for the cognitive test. NRCs were instructed to place all national items at the end of the questionnaires, and were told that the ISC would have to document and approve the items before they could be included in the final versions of the questionnaires.

System for translating student instruments

Translation of the student-level ICILS instruments into national languages were implemented in national centers using the AssessmentMaster© system (see Chapter 3 for more details about this system). The system was designed to (i) give the people within the national centers involved in the translation work (translators, translation reviewers, NRCs, international staff tasked with reviewing the materials) access to the assessment, and (ii) enable them to communicate with one another.

The information contained within each screen of the student materials visible to students consisted of a series of text elements that required translation. The text in each element varied in length from a single word through to a short paragraph. National centers were provided with the text in the source version and were required to translate or adapt this text as necessary. Different display modes allowed translators to enter text either with or without the use of the HTML tags embedded in the item material.

The system also allowed a full record of the editing history (saved changes) of each text element to be maintained. The system highlighted differences between the current

and previous versions and had facility to revert back to previous revisions. Users could preview each screen, with either the source language or the most recent saved changes to the test language on view, and they could switch between the two.

Users could also set the status of the translations for each screen, whether for internal purposes within national centers (e.g., Translating, Ready for Translation Review) or for the external verification processes conducted at the international level (e.g., Ready for Verification, Ready for Layout Verification).

International verification of the instruments

In addition to undergoing the internal review of translations carried out by each national center, all survey instruments went through a rigorous three-part international verification process: (i) adaptation review, (ii) translation verification, and (iii) layout verification. These three processes are described later in this chapter. International quality-control monitors also conducted an independent review of the translation verification record as part of the ICILS quality assurance program (see Chapter 9).

Documentation in national adaptations form

When translating and adapting the international version of the instruments for national use, national centers needed to make certain changes, selections, and adaptations to the survey instruments. In doing so, they were required to keep in mind that the objective of the study was to create an international database containing comparable data from all participating countries with complete documentation. Consequently, each change had to be recorded electronically on the national adaptations form (NAF). This form was used not only for documentation purposes but also when national data were added to the international database.

The NAF in Microsoft® Excel format consisted of several worksheets per cognitive test, student questionnaire, teacher questionnaire, ICT coordinator questionnaire, and principal questionnaire. Separate sheets for verifying the layout of paper instruments, online instruments, and codebooks were also supplied, as was information on the test language of the NAF and the stage of preparation.

The ISC asked national centers to complete a NAF for each survey language used in their country. The international center also asked national center staff to document whether they intended to include any national items and categories and, if so, to provide a description of their content in both the national language and English.

The NAFs were completed and reviewed at various stages of the verification process. During data-management seminars preceding the field trial and main survey, national center staff received detailed instructions on how to work with the NAFs and how to adapt the data entry software accordingly.

Adaptation review

ISC staff requested NRCs to consult with them during reviews of all proposed national adaptations. They 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 submitted the NAF for consultation with the ISC. After completing its review, the ISC provided the national centers with feedback on

their adaptations and, where necessary, suggested 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;
- 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 so that these updated forms (Version II) could inform the translation verification process.

Translation verification

International translation verifiers

The IEA Secretariat enlisted the assistance of an independent translation company, cApStAn Linguistic Quality Control (Brussels, Belgium), to verify the translations for each country. The international translation verifiers for ICILS needed to have the target language as their mother tongue, have formal credentials as translators working in English, be educated at university level, and (if possible) live and work in the target country (or be in close contact with it).

Verifiers attended a training seminar where they received detailed instructions for reviewing the survey instruments and registering deviations from the international version. They also received general information about the study and design of the instruments, together with a description of the translation procedures that the national centers used.

International translation verification

The primary tasks of the translation verifiers were to

- Evaluate the accuracy and the comparability of the national versions of the ICILS instruments to the international version in English;
- Document all deviations in the country's translation/adaptation; and
- Suggest alternative translations/adaptations that would improve the accuracy and comparability of the national versions.

Instructions given to verifiers emphasized the importance of maintaining the meaning and difficulty level of each test and questionnaire item. Specifically, verifiers had to ensure the following:

- The translation had not affected the meaning or reading level of the text;
- The test items had not been made easier or more difficult;
- No information had been omitted from or added to the translated text;
- The instruments contained all of the correct items and response options, in the same order as in the international version;
- All national adaptations implemented in the instruments were documented in the NAF.

The verifiers were required to work within the software platform developed for ICILS when making edits to and suggestions for the text of the student cognitive test and student questionnaire. The verifiers used the editing functions of Microsoft® Word ("Track Changes" and "Insert Comments") to document any errors or suggested changes directly in the teacher, principal, and ICT coordinator questionnaires. Verifiers were asked to provide suggestions that would improve the comparability of the instruments when appropriate, and to evaluate the overall quality, accuracy, and cultural relevance of the translation.

To help NRCs understand the comparability of the translated text with the international version, verifiers were asked to assign a "severity code" to any deviations. Descriptions of the degree of severity indicated by each code follow:

- **Code 1.** *Major change or error:* The translation affected the meaning in a way that meant respondents might not understand the text or could be misled by it. Examples included the incorrect order of choices in a multiple-choice item; incorrect order of items; omission or addition of a graphic, item, or answer option; incorrect translation resulting in the answer being suggested by the question; and an incorrect translation that changed the meaning or difficulty level of an item. In case of doubt, verifiers were instructed to use "Code 1?" so that the error could be referred to the ISC for further consultation.
- **Code 2**. *Minor change or error:* Examples included grammar mistakes and spelling errors that did not affect comprehension.
- **Code 3.** *Suggestion for alternative:* The translation was deemed adequate, but the verifier suggested a different wording (stylistic or fluency improvement).
- **Code 4.** *Acceptable change:* The change was deemed acceptable and appropriate, but was not necessarily documented in the NAF. An example of an acceptable adaptation is the case where a reference to winter was changed from January to July in the instruments for participating countries from the Southern Hemisphere.

On receiving the verification feedback, NRCs reviewed the translation verifiers' suggestions and revised the instruments according to this feedback. NRCs were also asked to complete a translation verification summary form after the field-trial verification, and to comment on verifier suggestions that they decided not to implement.

Results of the translation verification

The errors that the translation verifiers commonly found during the verification process included mistranslations, inconsistencies, omissions/additions, adaptations of names (fictional versus real), grammar, style (gender agreement, formality), and fluency issues ("free" versus "word-for-word" translations, Anglicisms). Some verifiers noted the challenge of translating and adapting certain ICT concepts and terms for the particular national context.

The extensive documentation collected via the NAF enabled verifiers to provide meaningful feedback on issues arising out of the national adaptations. When providing this feedback, verifiers took into account the matters that the national centers reported and the recommendations that the ISC made during the adaptation review.

The translation verifiers of the main survey instruments noted the great care with which their verification feedback from the field trial was implemented. Overall, the verifiers considered the national translations/adaptations to have been very well documented and

of high quality, thus making it possible to achieve a good balance between faithfulness and fluency.

Layout verification

Once translation verification had been completed, the ISC asked national centers to make any changes resulting from translation verification, compile their instruments, and notify the ISC that their materials were ready for layout verification. National center staff were asked to use the translation system to let the ISC know the *student instruments* were ready for this stage. They were asked to upload to a secure server the *teacher, principal, and ICT coordinator questionnaires* in PDF format for each test language in the main survey along with an updated NAF (Version III) containing any changes resulting from translation verification.

Staff at the ISC then accessed these files for layout verification. Two independent reviewers at the ISC reviewed each set of materials. They documented all layout issues identified in a worksheet added to the NAF. The layout issues in each set of instruments were grouped according to whether they 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 layout issues most commonly found in the student test and questionnaire were missing or additional line breaks, unrealistic URLs or email addresses, text not fitting within the predefined spaces, and issues with HTML tags. The verifiers often fixed, on behalf of the national center, technical issues relating to the translation software or, in some cases, referred them to the software developers.

The layout issues found in regard to the paper-based instruments were typically formatting ones (e.g., spacing, font size, margins, consistency across questions), incorrect order of questions, missing text, and the addition or omission of questions not agreed upon from the adaptation review.

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 asked to make the appropriate changes to their materials. They were also advised that there would be no 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.

Online/paper instrument verification

For countries administering the teacher, principal, and ICT coordinator questionnaires online, instrument preparation comprised an additional verification step. Countries were asked not to set up their online questionnaires until the paper-based instruments had been verified (as described above). Countries then used the IEA SurveySystem to set up the survey online, a process that was primarily a matter of copying and pasting text elements from the already verified paper instruments. The *Designer* component of the IEA SurveySystem enabled users to create, delete, disable, and edit survey components (e.g., questions and categories) and their properties. It allowed for translation of all text passages in the existing national paper questionnaires and additional system texts, and it included a complete web server able to verify and test-drive the survey exactly as if under live conditions. Once conversion was complete, the *Designer* also allowed users to export converted questionnaire files to the IEA DPC for final verification.

To ensure that data from both administration modes were comparable, the IEA DPC conducted a systematic check of the paper and online questionnaires. Except for a few inevitable exceptions, which were necessary because of the different administration modes and which were set down for NRCs in "online adaptation notes," any deviations with regard to content and layout between paper and online instruments were reported back to the countries. In such cases, NRCs were requested to update their online instruments to match the paper instruments.

As a final stage during production of the national online instruments, IEA DPC staff checked the layout and structure of all online questionnaires. They began by checking the online instruments against each national paper version that had undergone paper layout verification. This practice helped ensure that the instruments within one country were the same regardless of whether they would be administered on paper or online. DPC staff also conducted visual checks, using the same standards and procedures as those for verification of the paper layout.

Staff then checked the structure of the national online instruments against the structure of the international online instruments (e.g., number of categories and width of noncategorical questions). The only intended deviations they approved were those documented on the NAF.

All inconsistencies that were found were listed in the NAF and reported back to the NRCs for their review. Another staff member at the IEA DPC then verified the revised version of the instruments. This procedure was repeated until no more inconsistences could be found. For the majority of languages, one to two rounds were needed before the IEA DPC approved the layout and structure of the online instruments.

As a last check, the DPC set all instruments online and asked the NRCs to review the questionnaires one more time in their online environment. In a few cases, this check resulted in additional minor changes (e.g., correction of spelling errors). It was only after completion of this final check that respondents received notification that the questionnaires were ready and were given the link and login information they needed to access them.

Quality control monitor review

IEA hired international quality-control (IQC) monitors from each country to document the quality of the ICILS assessment administration, including the survey materials. An important part of the IQC monitors' responsibilities was that of carefully reviewing the instruments used during the main survey data collection. The IQC monitors compared the final version of the questionnaires and student cognitive tests against the translation verification record to ensure that the recommendations of the translation verifier had been appropriately addressed. More information on the IQC monitors' findings can be found in Chapter 9.

Summary

The survey instruments and verification procedures were developed through an extensive process of cooperation, independent review, and consensus. Detailed documents helped the national centers follow the internationally agreed procedures for preparing national instruments, and some additional quality-assurance measures were implemented to ensure international comparability. Reports from the verifiers indicated that the procedures for the translation and adaptation of the ICILS assessment and questionnaires were generally very well followed, and that the translated and adapted instruments were of high quality.

Reference

ICILS International Study Center. (2012). *IEA International Computer and Information Literacy Study (ICILS) 2013 survey operation procedures 3 for the main survey: Instrument preparation.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

CHAPTER 6:

Sampling Design and Implementation

Sabine Meinck

Introduction

International comparative surveys require samples to be drawn randomly to guarantee valid inferences from the observed data on the population features under study. Randomness is a key criterion to warrant 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, Rust, & Adams, 1999).

The ICILS sample design is referred to as a "complex" design because it involves multistage sampling, stratification, and cluster sampling. This chapter describes these features and gives details on target population definitions, design features, sample sizes, and achieved design efficiency. The focus is on presenting the international standard sampling design; the specific characteristics of each national sampling plan are given in Appendix C "Characteristics of National Samples."

The IEA Data Processing Center in Hamburg, Germany, in collaboration with the national research coordinators (NRCs) of each participating country or education system,[1] selected the school samples for ICILS. A series of manuals and the Within-School Sampling Software (WinW3S; IEA, 2012a, 2012b) supported NRCs in their sample activities. The sampling referee, Jean Dumais (Statistics Canada), gave advice on sampling methodology. He reviewed and adjudicated all national samples.

Target-population definitions

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

Definition: Students

ICILS defined the target population of students as follows:

> The student target population in ICILS consists of all students enrolled in the grade that represents eight years of schooling, counting from the first year of ISCED Level 1,[2] 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. Norway, however, decided to survey students and their teachers at the end of their ninth grade; the results for these countries were annotated accordingly. 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.

1 In the following, the term "country" is used when referring to countries or education systems.

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

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

Definition: Teachers

ICILS 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.[3]

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

- Any school staff who were attending to the needs of target-grade students but were not teaching any lessons (e.g., psychological counselors, chaplains);
- Assistant teachers and parent-helpers;
- Nonstaff teachers who were teaching (noncompulsory) subjects that were not part of the curriculum (e.g., cases where religion was not a regular subject and was being taught by external persons);
- Teachers who had joined a school after the official start of the school year.

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

Definition: Schools

In ICILS, schools were defined as follows:

> A school is one whole unit with a defined number of teachers and students, which can include different programs or tracks. The definition of "school" should be based on the environment that is shared by students, which is usually a shared faculty, set of buildings, social space and also often includes a shared administration and charter.
>
> Schools eligible for ICILS are those at which target grade students are enrolled.

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

- Subunits of larger "campus schools" (administrative "schools" consisting of smaller schools from different cities or regions) were to be regarded as separate schools for sampling purposes. If a part of a larger campus school was selected for ICILS, the principal/ICT coordinator of the combined school was asked to complete the school questionnaire with respect to the sampled subunit only.
- Schools consisting of two administrative units, but with shared staff, shared buildings, and offering some opportunities for the students to change from one school to the other, were to be regarded as one combined school for sampling purposes.

3 Teachers who were on long-term leave during the testing period (e.g., maternity or sabbatical leave) were not in scope of ICILS.

- The parts of a school with two or more different study programs that had different teaching staff, took place in different buildings, and offered no opportunity for students to change from one study program to the other were to 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.

Coverage and exclusions

Population coverage

The ICILS consortium encouraged all ICILS countries to include in the study all students and teachers included in the target population definition. However, countries could elect to remove larger groups of schools, students, or teachers from the target population for political, operational, or administrative reasons. This removal is referred to as *reduced national coverage*. No country chose this option; all provided full coverage of their target populations.

In most ICILS countries, smaller groups of schools, students, and teachers had to be removed from the target population for practical reasons, such as difficult test conditions or prohibitive survey costs. Such removals were regarded as *exclusions*. Some students and teachers were excluded because their entire school was excluded (*school-level exclusions*), while certain students were excluded within sampled and participating schools (*within-sample exclusions*).

It should be emphasized that the ICILS samples represented only the nationally defined target populations.

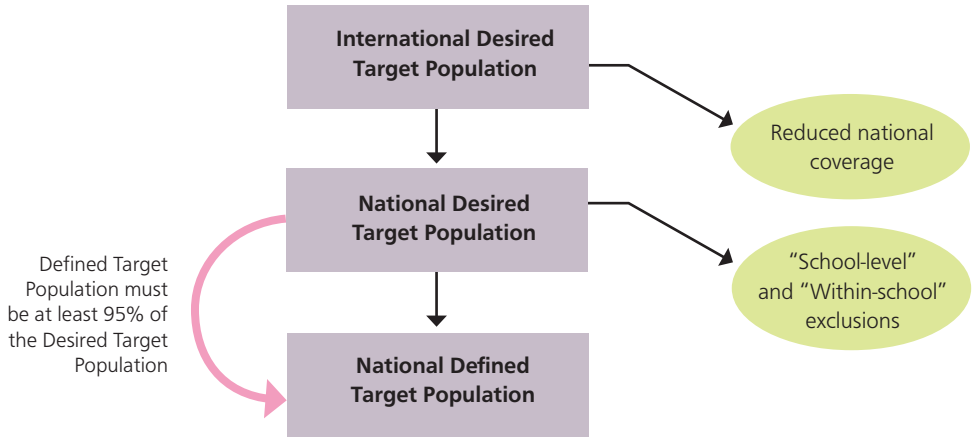
School exclusions

Table 6.1 gives an overview of the exclusion of schools and respective percentages for all participating countries. The percentages given are computed as the number of excluded schools divided by the total number of schools belonging to the national desired target population, multiplied by 100.

In most countries, very small schools and schools exclusively dedicated to 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 survey were only used to complement student- and teacher-level data, no specific thresholds were determined for exclusions at the school level.

School exclusions differed substantially across countries, a point that should be kept in mind when interpreting the results from the school-level data. Please note also that because school exclusions typically concerned small schools, the percentages of excluded schools always tended to be higher than the percentages of excluded students or teachers. Figure 6.1 summarizes the relationship between coverage and exclusions and the national defined target population.

Figure 6.1: Relationship between coverage and exclusions and the nationally defined target populations



Student exclusions

The overall exclusion rate of students is the sum of the students' school-level exclusion rate and the weighted within-sample exclusion rate. The school-level exclusions consisted of students excluded because their schools were excluded before the school sampling.

Unlike the method used to calculate the school exclusion rate described in the previous section, the student exclusion rate was calculated as the number of target-grade students in excluded schools divided by the total number of students belonging to the national desired target population, multiplied by 100. The respective figures were provided by the NRCs.

The within-sample exclusions were estimated on the basis of information collected from the sampled schools. The percentages given in Table 6.2 were then computed as the (estimated) total number of excluded students divided by the (estimated) total number of students belonging to the nationally desired target population, multiplied by one hundred.

Within-sample exclusions could consist of students with physical or mental disabilities or students who could not speak the language of the test (typically, students with less than one year of instruction in the test language). Any other type of within-sample student exclusions was not permitted.

Table 6.1 and Appendix C of this report provide details about the exclusion types for each country.

Each country was required to keep the overall proportion of excluded students (due to school-level and within-school exclusions) below five percent (after rounding) of the desired target population. In four of the countries participating in ICILS, the overall exclusion rate was above five percent, which resulted in respective annotations in the ICILS international report (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014).

Table 6.1: Percentages of schools excluded from the ICILS 2013 target population

Country	Excluded Schools (%)	Type of Exclusion
Australia	1.5	Special needs schools
	0.2	Schools instructing in a language and curriculum other than English
	0.1	Hospital schools, correctional schools
	0.4	Intensive English-language schools
	1.0	Alternative curriculum schools
	0.7	Geographically remote schools
	0.5	Distance education schools
	4.4	**Total**
Chile	5.6	Special needs schools
	2.7	Very small schools (fewer than five students)
	0.1	Geographically remote schools
	8.4	**Total**
Croatia	1.7	Students taught in language different from Croatian
	0.4	Different curriculum schools
	2.8	Special needs schools
	4.9	**Total**
Czech Republic	0.4	Schools with Polish as a language of instruction
	8.1	Special needs schools
	8.5	**Total**
Denmark	6.6	Day and also day and night treatment centers (*dagbehandlingstilbud og behandlingshjem*)
	2.4	Special schools for youth (*kommunale ungdomsskoler og ungdomskostskoler*)
	8.4	Special schools for children (*specialskoler for børn*)
	0.2	Unknown
	4.2	Very small schools (fewer than four students)
	21.7	**Total**
Germany	7.2	Special needs schools
	1.3	Waldorf schools
	8.5	**Total**
Hong Kong SAR	4.6	International schools
	11.3	Special schools
	7.2	Private schools
	23.2	**Total**
Korea, Republic of	5.4	Geographically remote schools
	0.3	Different curriculum schools: physical education middle schools
	0.3	Very small schools (fewer than five students)
	6.0	**Total**
Lithuania	0.1	Schools that are located in prisons
	5.4	Special needs schools
	4.0	Very small schools (fewer than seven students)
	9.5	**Total**
Netherlands	21.0	Special needs schools
	0.5	Very small schools (fewer than 10 students)
	0.1	International schools
	21.6	**Total**
Norway (Grade 9)	0.9	Schools with Sami language as medium of instruction
	5.9	Special needs schools

Table 6.1: Percentages of schools excluded from the ICILS 2013 target population (contd.)

Country	Excluded Schools (%)	Type of Exclusion
Norway (Grade 9)	1.9	Very small schools (fewer than five students)
	2.4	Steiner schools (ICT introduced in Grade 8)
	1.0	International schools
	12.0	**Total**
Poland	7.3	Special needs schools
	2.6	Very small schools (fewer than nine students)
	2.5	Other special schools
	12.4	**Total**
Russian Federation	*	Very small schools (fewer than four students)
	*	Special needs schools
	*	Evening schools (students under 18 years of age)
	*	**Total**
Slovak Republic	0.4	Students taught in other languages
	4.4	Special needs schools
	3.5	Very small schools (fewer than five students)
	8.3	**Total**
Slovenia	0.2	Waldorf schools
	0.4	Italian schools
	5.7	Special needs schools
	6.3	**Total**
Switzerland	22.3	Special needs schools
	5.8	Very small schools (fewer than five students)
	28.1	**Total**
Thailand	0.9	Very small schools (fewer than six students)
	0.3	Special needs schools
	1.3	**Total**
Turkey	1.6	Special needs schools
	0.1	Music and ballet schools
	0.8	Very small schools (fewer than six students)
	3.1	Geographically remote schools
	0.1	Private foreign schools
	5.6	**Total**
Benchmarking participants		
City of Buenos Aires, Argentina	1.8	Special needs schools
	1.8	**Total**
Newfoundland and Labrador, Canada	1.9	Native language of instruction
	1.9	**Total**
Ontario, Canada	7.6	Very small schools (fewer than seven students)
	7.6	**Total**

Note: *Statistics not available.

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

Country	Exclusions Prior to School Sampling (%)	Within-Sample Exclusions (%)	Overall School Exclusions (%)
Australia	0.7	4.3	5.0
Chile	2.8	1.7	4.5
Croatia	1.1	2.6	3.7
Czech Republic	1.0	0.6	1.7
Denmark	2.9	1.9	4.8
Germany	0.8	0.7	1.5
Hong Kong SAR	5.1	1.5	6.5
Korea, Republic of	0.8	0.5	1.3
Lithuania	2.8	1.5	4.3
Netherlands	2.9	1.9	4.7
Norway (Grade 9)	1.7	4.4	6.1
Poland	2.9	1.7	4.6
Russian Federation	2.9	3.0	5.9
Slovak Republic	2.6	2.6	5.1
Slovenia	1.3	1.1	2.3
Switzerland	2.2	1.8	3.9
Thailand	0.3	0.8	1.1
Turkey	2.0	1.2	3.2
Benchmarking participants			
City of Buenos Aires, Argentina	1.4	0.2	1.6
Newfoundland and Labrador, Canada	0.8	6.8	7.6
Ontario, Canada	0.6	4.4	5.0

Teacher 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 proportion of teachers in excluded schools. Because statistics about teachers per grade are rarely available, some countries could not provide exact figures. Instead, they provided very rough estimates or no estimates at all. Teacher exclusion rates exceeded five percent in Newfoundland and Labrador (Canada), Denmark, Hong Kong SAR, Lithuania, Poland, and the Slovak Republic. Statistics on the number of eligible ICILS teachers were not available for Australia, Buenos Aires (Argentina), Chile, Germany, Norway (Grade 9), the Russian Federation, and Turkey, which meant teacher exclusion rates could not be computed for these countries.

School sampling design

The IEA DPC used a stratified two-stage probability cluster sampling design in order to conduct the school sample selection for all ICILS countries. During the first stage of the sampling, schools were selected systematically with probabilities proportional to their size (PPS) as measured by the total number of enrolled target-grade students. During the second stage, the DPC used a systematic simple random sample approach to select students enrolled in the target grade within participating schools.[4]

4 A three-stage sampling design was implemented in the Russian Federation. Regions were sampled in the first stage. Schools and students were sampled in the second and third stages, respectively.

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

School sampling frame

In order to prepare the selection of school samples, national centers provided a comprehensive list of schools that included the numbers of students enrolled in the target grade. This list is referred to as the *school sampling frame*. To ensure that each ICILS 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 measure of size (MOS), which was usually the number of students enrolled in the target grade or an adjacent grade;
- Values for each of the intended stratification variables.

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) are geographic region, urbanization level, source of funding, and performance level. Generally, ICILS used stratification for the following reasons:

- To improve the efficiency of the sample design, thereby making survey estimates more reliable and reducing standard errors. (The national centers were asked to provide stratification variables that they expected would be closely associated with students' learning-outcome variables.)
- To apply disproportionate sample allocations to specific groups of schools.

The latter design feature was used if the country required estimates with high precision levels for specific subgroups of interest in the target population. For example, assume that a country wished to compare public and private schools but only 10 percent of the students in that country were attending private schools. In such a case, a proportional sample allocation would result in a sample containing an insufficient number of private schools to provide reliable estimates for this subpopulation.

Also, if the sample size for one of the subgroups was small, even larger differences between the two different school types might not appear as statistically significant. In this situation, an appropriately larger sample of private schools could be selected while keeping the original proportional sample allocation for public schools the same. This approach would achieve the required precision levels for subgroup comparisons as well as for national population estimates. In the example given above (split of 10% private and 90% public schools), the proportional sample allocation for a minimum total sample size of 150 schools would result in a sample of 15 private and 135 public schools, while oversampling would result in a larger school sample of, say, 50 private and (again) 135 public schools.

ICILS applied two different methods of stratification. *Implicit stratification* meant that the sampling frame was sorted, prior to sampling, by implicit stratification variables, thus providing a simple and straightforward method with which to achieve a fairly

proportional sample allocation across all strata. With explicit stratification, independent samples of schools were selected from each *explicit stratum*, and sample sizes for each explicit stratum were then assigned in order to achieve the desired sample precision overall and, where required, for subpopulations as well.

Each country applied different stratification schemes after discussing these 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 Explicit Strata	Implicit Stratification Variables
Australia	Public/private (2)	2	SES (3)
Chile	Schools with Grades 8 and 9/ Schools with Grade 8 only (2) Administration (3) Urbanization level (2)	11	Performance level: mathematics (4)
Croatia	Monthly income (4)	4	Gender (3); Finance type (3)
Czech Republic	School type (2) Region (6)	12	Region (14)
Denmark	None	1	Region (5); School type (3)
Germany	Track (gymnasium/nongymnasium/ special needs schools) (3) Federal state (Berlin/other federal states) (2)	5	School type within nongymnasium (4); Federal state (16)
Hong Kong SAR	Monthly income (4)	4	Gender (3); Finance type (3)
Korea, Republic of	Province (16)	16	School gender (3)
Lithuania	Language of instruction (2) Public/private (2) Urbanization level (2)	7	Language of instruction (8)
Netherlands	Provided track(s) at school (3)	3	None
Norway (Grade 9)	Performance level (4)	4	Language of instruction (2)
Poland	Creative school/Normal school (2) Score (3) Public/private (2)	7	Urbanization level (4)
Russian Federation	Region (43)	43	Urbanization level (2)
Slovak Republic	School type (2) Language of instruction (2)	4	Region (8)
Slovenia	Region (12)	12	Performance level: mathematics (3)
Switzerland	Cantons and regions (8)	8	Public/private (2); Language (3)
Thailand	Jurisdiction (5)	5	Region (5)
Turkey	Public/private (2)	2	Geographical region (7)
Benchmarking participants			
City of Buenos Aires, Argentina	Public/private (2)	2	Socioeconomic status (3)
Newfoundland and Labrador, Canada	Language of instruction (2)	2	None
Ontario, Canada	Language of instruction (3)	3	School type: funding (3); Region (6)

School sample selection

In order to select the school samples for the ICILS 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 nonresponse and disproportional sampling of subpopulations, this method would lead to "self-weighted" samples where all units sampled in the last stage of sampling have similar design weights. In reality, however, no single ICILS sample was self-weighting.

First, 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 were 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 12 out of the 21 ICILS countries were disproportionally allocated to explicit strata in order to achieve precise estimates for subgroups.[5]

School sample selection involved the following steps:

1. Splitting the school sampling frame by explicit strata. All following steps were done independently within each explicit stratum (if explicit stratification was used).
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 in the explicit stratum by the number of schools to be sampled from that stratum;
4. Determining a random starting point, a step that determines the first sampled school in the explicit stratum;
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 accumulated MOS was equal to or above the value for selection, the corresponding school was included in the sample.

Figure 6.2 visualizes the process of systematic PPS sampling within an explicit stratum. In this diagram, the schools in the sampling frame are sorted in descending order 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. Sampled schools are displayed in blue.

Joncas and Foy (2012) provide a more comprehensive description of the sampling process and use an illustrative example to do so.[6]

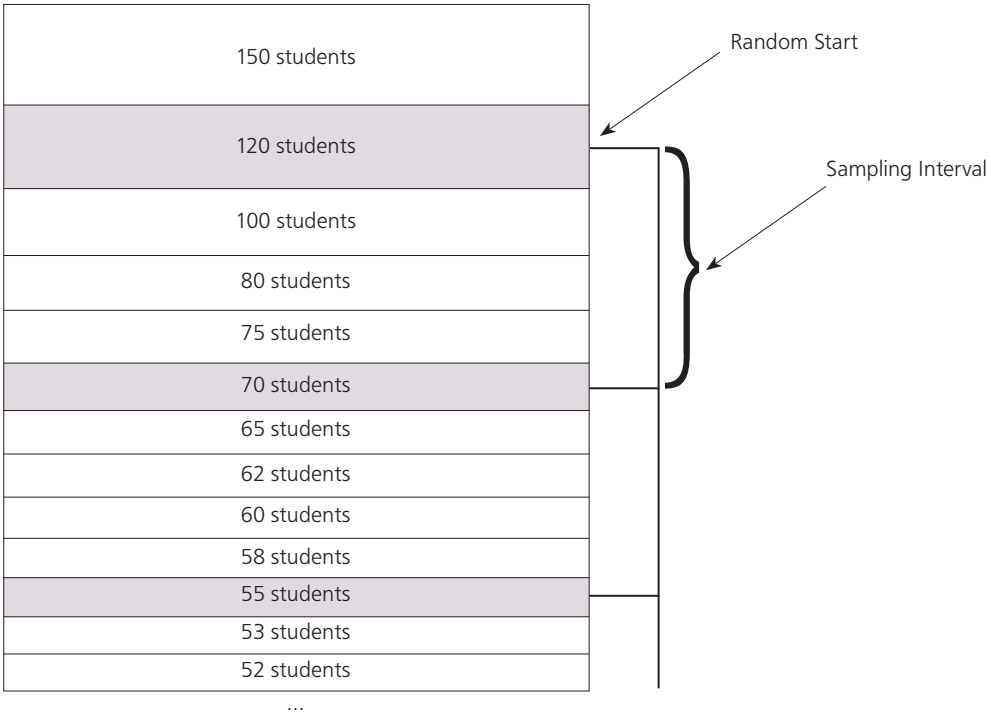
In some cases, the sampling design deviated from this general procedure:

- Very small schools were selected with equal selection probabilities to avoid large variations of sampling weights due to changing size measures. A school was regarded as "very small" if fewer students were enrolled in the target grade than in an average-size class in the explicit stratum.

5 See Chapter 7 for more details on sampling weights.

6 Access file "TIMSS and PIRLS Sampling Schools" in "Sample Design Details" at http://timssandpirls.bc.edu/methods/t-sample-design.html

Figure 6.2: Visualization of PPS systematic sampling



Source: Zuehlke (2011).

- Very large schools (i.e., schools with more students than the value of the sampling interval) were technically put into a separate stratum and selected with certainty (i.e., all schools in this category were included in the sample).

In order to reduce the considerable traveling costs of administering the study in the Russian Federation, the IEA DPC introduced another (first) sampling stage. The primary sampling units in this country were regions—83 in all. The 13 biggest regions were selected with certainty, and a further 30 regions were sampled from the remaining ones with PPS. During the second stage, the DPC selected an enlarged sample of 208 schools from the sampled regions in order to compensate for the increased sampling variance due to the additional sampling stage. The third sampling stage comprised, as in all other countries, random student sampling within the participating schools.

Most countries conducted a field trial with a small sample of schools one year before the main ICILS survey. In order to avoid response contamination and lower participation rates, the sampling team selected a school for either the field trial or the main survey whenever possible.

Because ICILS was conducted in the same year as the OECD Teaching and Learning International Survey (TALIS) 2013, several countries requested that schools not be selected for both studies. The IEA DPC collaborated closely with the TALIS sampling team to prevent school sample overlap whenever possible. The procedures used to prevent school overlap ensured randomness of selection and correct school selection probabilities for both studies.

Once schools had been selected from the sampling frame, up to two (nonsampled) replacement schools were assigned for each originally sampled school. The use of replacement schools in ICILS was limited (see Chapter 7 for more details). In order to

reduce the risk of nonresponse bias due to replacement, the replacement schools were typically assigned as follows: the school which appeared in the sorted sampling frame directly after a selected school was assigned as the selected school's first replacement, while the preceding school was used as its second replacement. This procedure ensured that replacement schools shared similar characteristics with the corresponding sampled schools to which they belonged, notably that they were of similar size and belonged to the same stratum.

Within-school sampling design

Within-school sampling constituted the second stage of the ICILS sampling process. The NRCs or their appointed data managers selected the students and teachers. The use of WinW3S software, developed by the IEA DPC, ensured the random selection of students and teachers within the sampled schools. Replacement of nonresponding individuals was not permitted.

Student sampling

WinW3S employed systematic stratified sampling with equal selection probabilities to select students from comprehensive lists of target-grade students provided by the participating schools. In order to ensure a nearly proportional allocation among subgroups and to increase sample precision, students were sorted by gender, class allocation, and birth year prior to sampling (implicit stratification).

In a few schools in the Netherlands and Switzerland, intact classrooms were sampled instead of single students.[7]

Teacher sampling

As occurred with the student sampling, 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, teachers were sorted by gender, main subject domain, and birth year prior to sampling (implicit stratification).

Sample size requirements

The ICILS consortium set, in line with practice in other IEA studies, high standards for sampling precision, and aimed to achieve reasonably small standard errors for survey estimates. The student sample needed to ensure a specified level of precision for population estimates, defined by confidence intervals of ±0.1 of a standard deviation for means, and ±5 percent for percentages. This requirement translated, with respect to ICILS' main outcome variable, that is, the students' computer and information literacy (CIL) score scale with a mean of 500 score points and a standard deviation of 100, into standard errors that needed to be below five scale score points.

The IEA DPC was responsible for determining sample sizes that were expected to meet these requirements for each participating country. With the exception of two

7 Classes were sampled with simple random sampling (SRS); they could not be replaced or substituted. However, national center staff could exclude a class from selection if it consisted solely of excluded students. All students from the selected class were asked to participate in the survey. This additional sampling step was considered at the stage when weights were being calculated. Chapter 7 provides the reasons behind this deviation from the international standard design.

participating education systems (Hong Kong SAR and Buenos Aires, Argentina) that failed to meet the IEA sample participation standards, all participating countries achieved this requirement (see Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014, p. 96). The required precision levels of percentages were also met for the vast majority of population estimates presented in the ICILS international report (Fraillon et al., 2014).

Other considerations also needed to be taken into account when determining the required number of students and teachers to sample:

- Some types of analysis, such as multilevel modeling, require a minimum number of valid cases at each sampling stage (see, for example, Meinck & Vandenplas, 2012);
- For the purpose of building scales and subscales, a minimum number of valid entries per response item is required;
- 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. Descriptions of these sample sizes follow.

School sample sizes

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

- Previous student surveys showed 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 fewer 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

In each sampled school, a minimum of 20 students in the target grade was randomly selected across all target-grade classes. In schools with fewer than 20 eligible students, all students were asked to participate in the survey. Also, if the number of eligible students was greater than 20 but fewer than or equal to 25, all students were selected to prevent smaller groups of nonselected students feeling excluded from the study.

Each country was required to have an achieved student sample size of about 3,000 tested students. Due to nonresponse, school closures, or other factors, some countries did not meet this requirement. The ICILS 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

In each sampled school, a minimum of 15 teachers was randomly selected for the survey.[8] For schools with fewer than 15 eligible teachers, all teachers were asked to participate in the survey. If the number of eligible teachers was greater than 15 but fewer than or equal to 20, all teachers were selected to prevent smaller groups of nonselected teachers feeling excluded. ICILS did not specify a minimum achieved teacher sample size.

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 could be treated as participants in the student survey but not in the teacher survey and vice versa due to specific minimum within-school response rate requirements. This requirement explains differences in the numbers of participating schools for the student and teacher surveys in most countries.[9]

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
Australia	325	311	5326	294	3495
Chile	180	174	3180	174	1800
Croatia	180	170	2850	179	2578
Czech Republic	170	170	3066	170	2126
Denmark	150	103	1767	82	728
Germany	150	136	2225	121	1386
Hong Kong SAR	150	118	2089	107	1338
Korea, Republic of	150	150	2888	150	2189
Lithuania	179	162	2756	163	2171
Netherlands	150	121	2197	96	1083
Norway (Grade 9)	150	138	2436	116	1158
Poland	158	156	2870	157	2228
Russian Federation	208	206	3626	207	2728
Slovak Republic	174	167	2994	167	2145
Slovenia	223	218	3740	214	2787
Switzerland	170	98	3225	74	796
Thailand	210	198	3646	184	2114
Turkey	150	141	2540	150	1887
Benchmarking participants					
City of Buenos Aires, Argentina	200	68	1076	49	591
Newfoundland and Labrador, Canada	155	118	1556	103	403
Ontario, Canada	202	193	3377	153	443

8 Due to special circumstances, the minimum sample size had to be reduced to five teachers per school in the two participating benchmarking entities of Canada (Newfoundland and Labrador and Ontario).

9 Please refer to Chapter 7 for details.

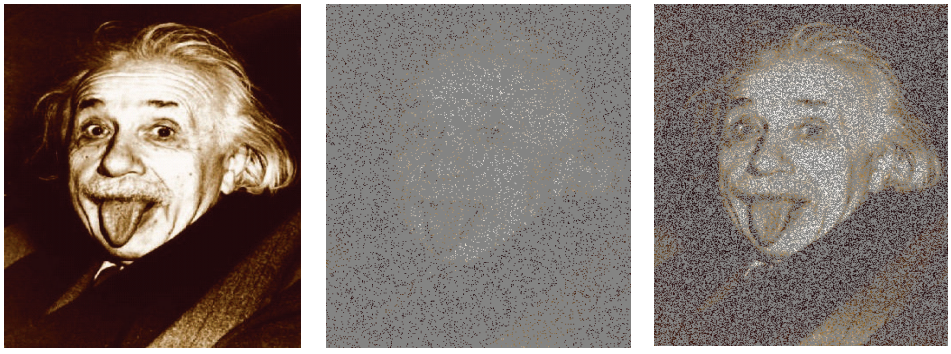
Efficiency of the ICILS sample design

As already noted, ICILS determined specific goals in terms of sampling precision, especially that standard errors should be kept below specific thresholds. Readers who are relatively unfamiliar with this topic may benefit from the following illustration of this concept.

In any sample-based survey, researchers want to use data collected from the sample in order 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 within specific design and cost limits. A measure of the precision is the standard error. As the standard error becomes larger, the picture becomes more "blurred" and inferences from sample data to populations less reliable. We can visualize this situation through reference to the following example.

Let us assume our population of interest is the left-hand picture in Figure 6.3 below—the famous picture of Einstein taken by Arthur Sasse in 1951. The picture consists of 340,000 pixels. We can draw samples from this picture, with each having more pixels than the one before it, and then reassemble the picture using only the sampled pixels. As can be seen in the middle of the right-hand picture of Figure 6.3, 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 6.3: 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 is 100, the population variance σ^2 will be 10,000, and the standard error of the estimated scale mean $\sigma_{\hat{\mu}}$ will equal five scale score points

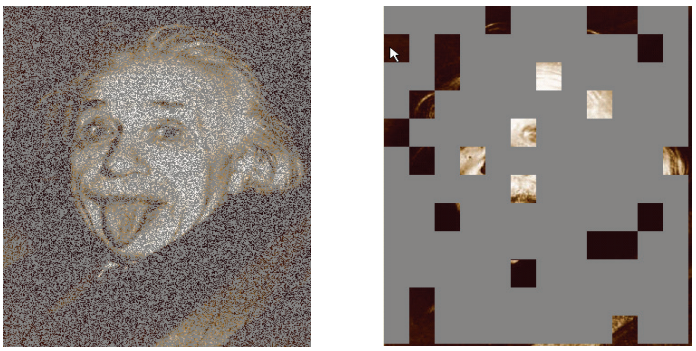
or fewer. Rearranging the formula above leads to a required minimum sample size of 400 students per country. As pointed out earlier, however, the actual minimum sample size for the countries participating in ICILS was 3,000 students.

The key reason for this sample size requirement is that ICILS did not employ SRS sampling but cluster sampling. Students in the sample were therefore members of "clusters" because groups of them belonged to the same schools.

Students within a school are more like one another than are students from different schools because they are exposed to the same environment and teachers; 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 being measured, applying this formula will most likely underestimate standard errors from cluster sample data. Figure 6.4 visualizes this effect through use 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 6.4: Sampling precision with equal sample sizes—simple random sampling versus cluster sampling



Note that stratification also has an influence on sampling precision. Choosing the stratification variables related to the outcome variables makes it possible to increase the sampling precision, unlike the situation with no nonstratified 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 provides another reason as to why the SRS formula for estimating sampling variance was not applicable for the ICILS 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 that should be used to correctly estimate standard errors for ICILS data.

The tables below present the achieved efficiency of the ICILS sampling design, which is measured by the design effect as

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

Here, 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 database but where the sample is a simple random sample (with replacement, conditional on the achieved sample size of the variable of interest).[10] 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 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 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 had been 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 stratification and clustering design.

Tables 6.5 and 6.6 provide the design effects of the ICILS main outcome variables for students and for teachers in each participating country. This information helps determine sample sizes and design strategies in future surveys with similar objectives. The design effects vary for different scales, but this is not unusual because the similarity of students and teachers within schools should be higher when the survey instruments ask about school-related matters than, for example, how frequently they (teachers and students) use ICT as individuals. The last column in both tables gives the effective sample sizes.

In Table 6.5, we can see that the effective sample size relates to the design effect of the CIL 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, the national samples of all countries except Hong Kong SAR, the Netherlands, Switzerland, and the City of Buenos Aires (Argentina), none of which met IEA sample participation requirements, achieved or exceeded the envisaged effective sample size of 400. Table 6.6 shows that most countries estimated the effective teacher sample sizes of 400 or above.

The average design effect of the CIL scale of 4.7 is smaller than in other studies with comparable sampling designs, such as the Programme for International Student Achievement (PISA [OECD, 2012]).[11] We can take this as a sign that CIL is not subject to cluster effects as large as, for example, mathematics achievement. Most other scales pertaining to the student survey showed even lower design effects. The teacher-related scales had an average design effect of 2.8 across all countries.

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

11 Because of its age-based population definition, PISA samples include students from different grades and can therefore be expected to have even lower clustering effects than ICILS if the same subject domain is measured.

Table 6.5: Design effects of main outcome variables, student survey

Country	Design Effects Using Scale Scores Pertaining to Students' Use of and Engagement with ICT at Home and School										Computer and Information Literacy: Plausible Values 1–5		
	S_ADVEFF	S_BASEFF	S_TSKLRN	S_USEAPP	S_USELRN	S_USEREC	S_USESTD	S_USECOM	S_INTRST	S_USEINF	Design effect	Sample size	Effective sample size
Australia	1.8	1.8	3.0	2.9	10.4	2.4	4.1	1.4	2.1	1.8	4.0	5326	1324
Chile	2.3	2.2	2.7	2.3	6.3	2.1	2.0	2.2	3.2	2.5	4.0	3180	804
Croatia	1.5	2.5	1.6	1.5	2.9	1.2	1.1	1.8	1.1	1.3	2.4	2850	1192
Czech Republic	2.0	1.3	2.7	2.1	3.4	1.3	2.3	1.9	1.8	1.5	2.8	3066	1099
Denmark	1.2	1.1	1.4	2.1	6.0	1.5	3.0	1.9	1.6	1.0	2.8	1767	630
Germany	1.6	1.7	1.9	2.4	2.6	1.3	1.7	1.4	1.4	1.3	1.9	2225	1199
Hong Kong SAR	1.7	3.1	3.3	2.5	3.4	1.3	3.5	1.6	2.2	1.0	12.3	2089	170
Korea, Republic of	1.1	1.4	1.6	1.4	1.5	1.4	1.9	1.5	1.8	0.8	2.4	2888	1227
Lithuania	1.5	1.8	2.4	2.6	3.2	1.4	3.2	1.2	1.2	1.6	4.4	2756	631
Netherlands	1.4	2.6	2.2	2.2	7.4	1.6	4.0	1.9	1.9	1.7	6.2	2197	352
Norway (Grade 9)	1.1	1.3	1.8	1.7	2.8	1.4	2.7	1.3	1.4	1.3	2.3	2436	1048
Poland	1.7	2.1	1.6	1.7	3.2	1.3	1.3	1.1	1.8	1.9	2.5	2870	1167
Russian Federation	2.2	1.6	3.0	3.3	3.1	1.9	2.9	2.3	1.3	1.8	3.3	3626	1102
Slovak Republic	1.4	1.2	1.7	1.1	1.9	1.4	1.1	1.0	1.3	1.1	6.3	2994	473
Slovenia	2.5	1.4	2.8	2.3	3.7	1.7	2.3	2.4	2.2	2.4	3.4	3740	1095
Switzerland	5.3	4.6	5.5	4.5	7.3	4.8	7.0	5.2	4.2	6.2	9.6	3225	337
Thailand	3.4	4.0	4.4	3.6	3.5	3.3	2.9	5.1	3.5	4.5	7.3	3646	501
Turkey	2.4	2.3	2.7	2.0	3.0	2.4	2.0	2.6	2.2	1.7	5.8	2540	441
Benchmarking participants													
City of Buenos Aires, Argentina	1.5	2.5	2.4	1.9	6.0	0.9	2.8	1.0	1.4	2.2	8.7	1076	123
Newfoundland and Labrador, Canada	0.9	2.1	0.7	1.5	2.0	1.6	1.8	1.5	1.2	1.4	1.6	1556	947
Ontario, Canada	2.0	2.5	3.1	2.7	5.4	2.7	3.1	2.1	2.7	3.6	5.7	3377	588
Average	**1.9**	**2.2**	**2.5**	**2.3**	**4.2**	**1.8**	**2.7**	**2.0**	**2.0**	**2.0**	**4.7**	**2830**	**783**

Key:

Scale scores

S_ADVEFF — ICT self-efficacy advanced skills
S_BASEFF — ICT self-efficacy basic skills
S_TSKLRN — Learning of ICT tasks at school
S_USEAPP — Use of specific ICT applications
S_USELRN — Use of ICT during lessons at school
S_USEREC — Use of ICT for recreation
S_USESTD — Use of ICT for study purposes
S_USECOM — Use of ICT for social communication
S_INTRST — Interest and enjoyment in using ICT
S_USEINF — Use of ICT for exchanging information

Table 6.6: Design effects of main outcome variables, teacher survey

Country	Design Effects Using Scale Scores Pertaining to Teachers' Use of and Engagement with ICT at Home and School								Mean of Scale Scores		
	T_USEAPP	T_EFF	T_USELRN	T_USETCH	T_EMPH	T_VWPOS	T_VWNEG	T_COLICT	Design effect	Sample size	Effective sample size
Australia	2.5	2.2	2.4	2.7	2.2	2.8	3.7	7.9	3.3	3495	1058
Chile	3.2	3.3	4.3	5.1	4.6	2.9	3.4	3.5	3.8	1800	476
Croatia	1.6	1.8	1.5	1.5	1.8	2.0	1.6	1.7	1.7	2578	1532
Czech Republic	2.8	1.8	2.4	2.8	2.1	2.3	2.3	3.8	2.5	2126	836
Denmark	2.5	2.9	1.8	2.6	2.0	1.2	2.5	1.8	2.2	728	335
Germany	2.1	1.8	2.2	2.2	1.7	1.2	1.5	3.4	2.0	1386	690
Hong Kong SAR	1.6	1.2	2.1	2.3	1.8	1.4	1.9	1.8	1.8	1338	753
Korea, Republic of	2.3	2.6	3.5	2.3	1.5	2.5	1.7	2.4	2.4	2189	931
Lithuania	1.8	2.1	2.1	2.4	1.8	1.3	2.1	2.2	2.0	2171	1098
Netherlands	0.9	0.8	1.5	1.2	1.4	2.2	1.1	2.3	1.4	1083	755
Norway (Grade 9)	1.4	2.2	1.9	2.2	1.7	1.7	2.1	1.9	1.9	1158	613
Poland	2.6	2.5	2.3	2.3	2.5	2.1	2.7	2.7	2.5	2228	903
Russian Federation	2.8	4.4	2.4	3.2	2.6	4.7	6.2	3.9	3.8	2728	720
Slovak Republic	2.5	0.9	2.8	2.7	2.5	3.0	2.8	3.2	2.5	2145	842
Slovenia	2.7	2.8	2.1	2.9	2.0	2.5	2.4	4.0	2.7	2787	1043
Thailand	5.3	5.6	6.4	6.9	3.7	10.2	19.7	17.5	9.4	2114	225
Turkey	4.4	4.1	4.5	4.9	3.9	2.3	2.2	6.3	4.1	1887	464
Benchmarking participants											
Newfoundland and Labrador, Canada	0.7	1.2	1.0	1.1	1.2	1.9	1.7	1.0	1.2	403	330
Ontario, Canada	1.9	1.7	2.0	1.9	2.1	3.3	2.1	2.4	2.2	443	203
Average	**2.4**	**2.4**	**2.6**	**2.8**	**2.3**	**2.7**	**3.3**	**3.9**	**2.8**	**1831**	**727**

Key:

Scale scores

T_USEAPP — Use of specific ICT applications
T_EFF — ICT self-efficacy
T_USELRN — Use of ICT for learning at school
T_USETCH — Use of ICT for teaching at school
T_EMPH — Emphasis on teaching ICT skills
T_VWPOS — Positive views on using ICT in teaching and learning
T_VWNEG — Negative views on using ICT in teaching and learning
T_COLICT — Collaboration between teachers in using ICT

Summary

The ICILS student target population consisted of students enrolled in the grade that represented eight years of schooling (counted from the first year of primary school, that is, ISCED 1), providing that the students' mean age at the time of testing was at least 13.5 years. The teacher target population consisted of teachers teaching regular school subjects to students of the target grade.

The international sample design used for ICILS was a stratified two-stage probability cluster design. During the first stage, schools were sampled with probabilities proportional to their size. During the second stage, stratified systematic random sampling with equal selection probabilities was used to select target-grade students and teachers within participating schools. ICILS required, as a default, a minimum sample size of 150 schools, in which 20 students and 15 teachers were selected for the study. The national samples were designed to yield a student sample size of roughly 3,000 tested students.

National centers were allowed to exclude groups of students from the study for practical reasons. However, each country was required to keep the overall rate of excluded students below five percent of the target population. Only four countries slightly exceeded this maximum exclusion rate.

References

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*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA Data Processing and Research Center. (2012a). *Windows® Within-School Sampling Software (WinW3S)* [Computer software]. Hamburg, Germany: Author.

IEA Data Processing and Research Center. (2012b). *International Computer and Information Literacy Study: Survey operation procedures Unit 4, main survey*. Data collection and quality monitoring procedures. Hamburg, Germany: Author.

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. http://timss.bc.edu/methods/t-sample-design.html

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, 5–172.

Organisation for Economic Co-operation and Development (OECD). (2012). *PISA 2009 technical report*. Paris, France: OECD Publishing. http://dx.doi.org/10.1787/ 9789264167872-en

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

Zuehlke, O. (2011). Sample design and implementation. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS technical report* (pp. 59–68). Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

CHAPTER 7:

Sampling Weights, Nonresponse Adjustments, and Participation Rates

Sabine Meinck and Diego Cortes

Introduction

One of ICILS' major objectives was to obtain estimates of population characteristics. This objective implies that the study did not intend merely to describe sampled units but also to draw inferences on population properties. In order to draw correct conclusions about population features of interest, the ICILS researchers needed to take into account the complex sample design implemented in all participating countries (see Chapter 6 for details). In contrast to features in simple random samples (SRS), features in a complex sample may not be unbiased estimates of the corresponding population features, which is why complex samples must not be treated like SRS during data analysis.

A critical characteristic of a complex sample is that sampling units do not have equal selection probabilities. In ICILS, this characteristic applied to the sampled students, teachers, and schools. Furthermore, nonparticipation has the potential to bias results, and differential patterns of nonresponse increase this risk. To account for these complexities, the IEA research team computed sampling weights and nonresponse adjustments within each participating country, a process which led to an estimation (or "final") weight for each sampled unit.[1] All findings presented in ICILS reports are based on weighted data. Anyone conducting secondary analysis of ICILS data should follow this approach.

This chapter first describes the conditions under which students, teachers, and schools were deemed to be "participants." Descriptions of how the several sets of weights and nonresponse adjustments were computed follow. Please note that Chapter 13 of this report covers the creation of the replicate weights (needed for variance estimation). Subsequent sections describe the computation of participation rates at each sampling stage, the minimum participation requirements, and the achieved quality of sample implementation for each country. The ICILS research team regarded response rates as an important indicator of data quality. The final section of this chapter presents the results of analyses of the nonresponse patterns.

Types of sampling weights

The ICILS 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.

1 For further reading on the topic, we recommend Franklin and Walker (2003), Groves, Fowler, Couper, Lepkowski, Singer, and Tourangeau (2004), Meinck (2015), and Rust (2014).

- *Nonresponse adjustments:* The aim of these adjustments is to compensate for the potential for bias due to nonparticipation of sampled units. As with base weights, nonresponse adjustments are computed separately for each sampling stage. The main principle at work is that the (base) weight of the nonrespondents within a specific adjustment cell must be redistributed among the responding units in that cell. This "adjustment cell" contains sampling units that share specific features. For example, all private schools in a given region could comprise a stratum of schools, and it is from this stratum that a sample of schools would be 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 nonparticipating schools. This approach allows us to exploit the (usually) little information we have available about respondents and nonrespondents, and to assume that school nonparticipation is associated with the different strata (see also Lohr, 1999). The approach also assumes a noninformative response model, thus implying that nonresponse occurs completely at random within the adjustment cell (i.e., in ICILS, within a stratum).

Calculating student weights

School base weight (*WGTFAC1*)

The first sampling stage typically involved selecting schools in each country; the school base weight reflects the selection probabilities of this sampling step. Explicit stratification saw the school samples selected independently from within each explicit stratum h, with $h = 1,\ldots, H$. If no explicit strata were formed, ICILS regarded the entire country as one explicit stratum.

Systematic random samples of schools were drawn in all countries, with the selection probability of school i being 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 some 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.

In the Russian Federation, the first sampling stage involved selection of regions. Therefore, each school weight was multiplied by a region weight component that reflected the probability of selecting that region.

School nonresponse adjustment (*WGTADJ1S*)

Given the fact that some schools refused to participate in ICILS or had to be removed from the international dataset due to low within-school participation, the school base weights for participating schools had to be adjusted to account for the loss in sample size. Adjustments were calculated within nonresponse groups defined by the explicit strata. A school nonresponse 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, because $n_h^{s,e}$ was restricted to schools deemed eligible in ICILS. 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, perhaps because they had recently closed, did not have students in the target grade, or had only excluded students enrolled. Ineligible schools such as these were not taken into account during calculation of the nonresponse adjustment.

Class base weight (*WGTFAC2S*)

Some countries faced specific challenges when endeavoring to secure school participation in ICILS. In particular, the fact that ICILS sampled students across all target grade classrooms—in contrast to selecting specific ones—made it difficult to persuade some schools to participate in the survey. This issue, evident in a small number of schools in the Netherlands and Switzerland, was addressed through special arrangements wherein simple random sampling was used to select classes during the second sampling stage. In all affected schools, each student's weight was multiplied by a class weight component that reflected the selection probability of that class.

For each sampled class j, the class base weight was given by

$$WGTFAC2S_{hij} = \frac{C_{hi}}{c_{hi}^s}$$

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

Because relatively few schools were affected, *WGTFAC2S* is not part of the public use data files but instead is included in the *student base weight* (*WGTFAC3S*; see next section). Hence, the *student base weight* reflects, for all students in the database, the within-school selection probabilities, including those schools where class sampling was applied.

Student base weight (*WGTFAC3S*)

The standard within-school sampling approach in the student survey involved using the software WinW3S during the second sampling stage (IEA Data Processing and Research Center, 2012; see Chapter 8 for details) in order to conduct a systematic random selection of students. 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 and m_{hi} is the number of sampled students in school *i* in stratum *h*.

In those schools in the Netherlands and Switzerland where sampling involved classes, all eligible students in the selected classes were automatically selected for the survey. The student base weight consequently equaled one for these students. Similarly, in schools with fewer than 26 target-grade students, all eligible students were selected for participation. In these cases, the weight factor was also set to a value of one.

Student nonresponse adjustment (*WGTADJ3S*)

Unfortunately, not all selected students were able or willing to participate in ICILS. To account for the reduction in sample size due to within-school nonparticipation, the ICILS research team introduced a nonresponse adjustment factor. Given the lack of information about absentees, nonparticipation had to be assumed, for weighting purposes, as being completely at random within schools. This meant that participating students represented both participating and nonparticipating students within a surveyed school. It also meant that the sampling weights for these students had to be adjusted accordingly.

The adjustment for student nonresponse 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 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.

The weight was also adjusted for students who did not participate in the survey because they had left the sampled school after the within-school sampling. These students were assumed to remain part of the target population (they moved to a different school but had a zero chance of selection because within-school sampling was already completed). Excluded students within participating schools carried their weight (i.e., represented such students in the population) and therefore contributed to the overall estimates of exclusion.

Final student weight (*TOTWGTS*)

The final student weight of student *k* in school *i* in stratum *h* was the product of the four student-weight components:

$$TOTWGTS_{hik} = WGTFAC1_{hi} \times WGTADJ1S_{hi} \times WGTFAC3S_{hik} \times WGTADJ3S_{hik}$$

In schools with class sampling, the final student weight of student *k* in school *i* in stratum *h* was the product of the four following components:

$$TOTWGTS_{hik} = WGTFAC1_{hi} \times WGTADJ1S_{hi} \times WGTFAC2S_{hik} \times WGTADJ3S_{hik}$$

Calculating teacher weights

School base weight (*WGTFAC1*)

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

School nonresponse adjustment (*WGTADJ1T*)

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

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

Here, $n_h^{s,e}$ is again the number of sampled eligible schools and $n_h^{p\text{-}tch}$ 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 software 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 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, the weight factor was one.

Teacher nonresponse adjustment (*WGTADJ2T*)

Not all teachers were willing or able to participate. Therefore, participating teachers represented both participants and nonparticipants. Again, the nonresponse adjustment carried out within a given school assumed, for weighting purposes, that a random

process underlay the teachers' participation. For more details on this topic, see the last section (headed "Nonresponse Analysis") of this chapter.

The nonresponse 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 before the data collection were regarded as out of scope; weights were not adjusted in these instances.

Teacher multiplicity factor (*WGTFAC3T*)

Some teachers in ICILS were teaching at the target grade in more than one school and therefore had a larger selection probability. In order to account for this, a "teacher multiplicity factor" was calculated (see immediately below) as the inverse of the number of schools in which the teacher was teaching:[2]

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

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* was 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 was designed as a student and teacher survey, and therefore not specifically as a school survey. However, in order to collect background information at school level, the ICILS consortium handed out a principal questionnaire and an ICT coordinator questionnaire to every participating school. School weights were calculated and included in the international database in order to allow for analyses at the school level. Statements about school-level variables have to be treated with some caution, though, as they may be subject to large sampling errors.

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 that did not complete any item in either the principal questionnaire or the ICT coordinator questionnaire were regarded as nonparticipants in the school survey. In order to account for these nonresponding schools, the ICILS team calculated a school weight adjustment component for each participating school *i* as follows:

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

2 This information was taken from the teacher questionnaire.

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

Note that some schools may have been nonparticipants in the school survey but still have produced student and/or the teacher survey data. Consequently, some schools could be regarded as participants in the student and/or the teacher survey but nonparticipants in the school survey. Alternatively, some schools that completed (at least one of) the school-level questionnaires were regarded as nonparticipating in the student and/or teacher survey. It is vitally important that anyone conducting analyses of data from different file types keeps these points in mind. During this kind of multivariate analyses, missing values tend to accumulate as the number of variables increases. Individuals conducting secondary analyses should therefore closely monitor the potential loss of information due to missing data.

Final school weight (*TOTWGTC*)

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

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

Calculating participation rates

During ICILS, weighted and unweighted participation rates were calculated at student and teacher levels to facilitate the evaluation of data quality and lessen the risk of potential biases due to nonresponse. In contrast to the weight-adjustment processes described earlier, participation rates were computed first for the *originally sampled* schools and then for the replacement schools as well.

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 nonparticipating schools in the student survey. Let n^{op}, n^{fp}, and n^{np} denote the number of schools in each of the respective sets. The unweighted school participation rate in the student survey *before* replacement is then calculated as

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

and 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 m^{fp} be the set of eligible and participating students in all participating schools, that is, in the schools constituting *fp*, the full set of eligible participating schools. Let m^{np} be the set of eligible but nonparticipating students in schools constituting *fp*, and let m^{sfp} and m^{snp} be the number of students in the respective groups. The unweighted student response rate is then computed as

$$UPRS_{students} = \frac{m^{sfp}}{m^{sfp} + m^{snp}}$$

Note that the ICILS team did not consider it necessary to compute student response rates separately for originally sampled and replacement schools because the nonresponse patterns did not vary between (the participating) originally sampled and replacement schools.

The unweighted overall participation rate in the student survey *before* replacement then becomes

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

and the unweighted overall participation rate in the student survey *after* replacement becomes

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

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 was 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 again as follows, but only after the replacement schools had been taken into account:

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

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

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

and the weighted overall participation rate in the student survey *after* replacement was therefore

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

The only schools treated as participants in the student survey were those which had a participation rate of at least 50 percent of their sampled students. A school that did not meet this requirement was regarded as a nonparticipating school in the student

survey. Although the nonparticipation of this school affected the school participation rate, the students from this school were not included in the calculation of the student participation rate.

Overview of participation rates in the student survey

Tables 7.1 and 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, Switzerland's unweighted school participation rate was markedly higher than the weighted rate because almost all participating schools were in small cantons (where oversampling occurred), and only a very small number were in large cantons (where oversampling did not occur).

Table 7.1: Unweighted participation rates, student survey

Country	School Participation Rate		Student Participation Rate (%)	Overall Participation Rate	
	Before replacement (%)	After replacement (%)		Before replacement (%)	After replacement (%)
Australia	95.1	96.0	86.8	82.5	83.3
Chile	93.7	100.0	93.2	87.3	93.2
Croatia	94.4	94.4	85.3	80.5	80.5
Czech Republic	99.4	100.0	93.7	93.2	93.7
Denmark	38.3	73.0	87.7	33.6	64.1
Germany	72.5	91.3	82.1	59.5	75.0
Hong Kong SAR	74.0	78.7	88.7	65.6	69.8
Korea, Republic of	100.0	100.0	96.5	96.5	96.5
Lithuania	87.9	93.1	92.2	81.1	85.9
Netherlands	50.0	81.8	88.0	44.0	72.0
Norway (Grade 9)	85.2	92.6	90.0	76.7	83.4
Poland	84.8	98.7	85.5	72.5	84.4
Russian Federation	99.0	99.0	93.4	92.5	92.5
Slovak Republic	92.3	98.8	93.2	86.0	92.1
Slovenia	93.3	97.8	92.0	85.8	89.9
Switzerland	43.7	58.7	91.5	40.0	53.7
Thailand	88.0	94.7	93.7	82.5	88.8
Turkey	93.3	94.0	91.3	85.2	85.8
Benchmarking participants					
City of Buenos Aires, Argentina	68.0	68.0	80.3	54.6	54.6
Newfoundland and Labrador, Canada	98.3	98.3	88.0	86.5	86.5
Ontario, Canada	95.5	97.0	92.4	88.3	89.7

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 (%)
Australia	97.5	98.0	88.1	85.9	86.3
Chile	94.8	100.0	93.4	88.5	93.4
Croatia	94.7	94.7	85.6	81.1	81.1
Czech Republic	99.5	100.0	93.7	93.3	93.7
Denmark	41.8	73.0	87.8	36.7	64.1
Germany	70.9	91.3	82.4	58.4	75.2
Hong Kong SAR	72.4	77.0	89.1	64.5	68.6
Korea, Republic of	100.0	100.0	96.3	96.3	96.3
Lithuania	90.9	96.6	92.0	83.6	88.8
Netherlands	50.1	81.9	87.7	44.0	71.9
Norway (Grade 9)	84.8	92.8	89.8	76.2	83.4
Poland	84.7	99.3	87.0	73.6	86.3
Russian Federation	99.2	99.2	93.6	92.8	92.8
Slovak Republic	94.9	99.6	92.7	87.9	92.3
Slovenia	90.7	98.4	91.5	83.0	90.0
Switzerland	30.3	48.5	89.7	27.2	43.5
Thailand	89.5	94.9	93.6	83.8	88.8
Turkey	93.3	93.9	91.4	85.2	85.8
Benchmarking participants					
City of Buenos Aires, Argentina	67.5	67.5	80.2	54.2	54.2
Newfoundland and Labrador, Canada	98.3	98.3	87.8	86.3	86.3
Ontario, Canada	94.5	96.7	92.1	87.0	89.1

Unweighted participation rates in the teacher survey

The computation of participation rates in the teacher survey followed the same logic as that 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 then computed as

$$UPRT_{schools\_BR} = \frac{n^{op}}{n^{fp} + n^{np}}$$

and 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 t^{fp} be the set of eligible and participating teachers in schools that constitute *fp*, let t^{np} be the set of eligible but nonparticipating teachers in schools that constitute *fp*, and let t^{tfp} and t^{tnp} be the number of teachers in the respective groups. The unweighted teacher response rate can then be defined as

$$UPRT_{teachers} = \frac{t^{tfp}}{t^{tfp} + t^{tnp}}$$

Note that ICILS deemed it unnecessary to compute teacher response rates separately for the (participating) originally sampled and replacement schools because the nonresponse patterns did not vary between the sampled and replacement schools.

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

$$UPRT_{overall\_BR} = UPRT_{schools\_BR} \times UPRT_{teachers}$$

and the unweighted overall participation rate in the teacher survey after replacement was 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 was calculated as

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

while the weighted school participation rate in the teacher survey *after* replacement was calculated as

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

The weighted teacher participation rate was therefore

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

while the weighted overall participation rate in the teacher survey *before* replacement was

$$WPRT_{overall\_BR} = WPRT_{schools\_BR} \times WPRT_{teachers}$$

and the weighted overall participation rate in the teacher survey *after* replacement was

$$WPRT_{overall\_AR} = WPRT_{schools\_AR} \times WPRT_{teachers}$$

Note that the only schools counted as participants in the teacher survey were those where at least 50 percent of their sampled teachers had completed the survey. A school that did not meet this requirement was regarded as a nonparticipating school in the teacher survey. Although the nonparticipation of this school had an effect on the school participation rate, the teachers from this school were not included in the calculation of the teacher participation rate.

Overview of participation rates in the teacher survey

Tables 7.3 and 7.4 display the unweighted and weighted participation rates of all countries in the teacher survey. The discrepancies between the two tables again indicate differential response patterns between strata with disproportional sample size allocations. As before, Switzerland provides a very prominent example of this effect.

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 (%)
Australia	89.8	90.7	86.1	77.4	78.2
Chile	93.7	100.0	96.0	89.9	96.0
Croatia	99.4	99.4	96.6	96.1	96.1
Czech Republic	99.4	100.0	99.9	99.3	99.9
Denmark	31.9	58.2	85.0	27.1	49.5
Germany	66.4	81.2	79.7	53.0	64.8
Hong Kong SAR	66.7	71.3	82.6	55.1	59.0
Korea, Republic of	100.0	100.0	99.9	99.9	99.9
Lithuania	88.5	93.7	89.0	78.8	83.4
Netherlands	39.2	64.9	77.1	30.2	50.0
Norway (Grade 9)	71.8	77.9	83.7	60.1	65.2
Poland	85.4	99.4	94.1	80.4	93.5
Russian Federation	99.5	99.5	98.3	97.8	97.8
Slovak Republic	92.3	98.8	97.8	90.3	96.7
Slovenia	91.9	96.0	93.4	85.9	89.6
Switzerland	31.7	44.3	73.6	23.3	32.6
Thailand	81.3	88.0	97.4	79.2	85.7
Turkey	99.3	100.0	96.0	95.4	96.0
Benchmarking participants					
City of Buenos Aires, Argentina	49.0	49.0	81.2	39.8	39.8
Newfoundland and Labrador, Canada	85.8	85.8	91.4	78.4	78.4
Ontario, Canada	76.6	77.7	92.5	70.9	71.8

ICILS standards for sampling participation

Despite the efforts of each ICILS country to achieve full participation (i.e., 100%), high levels of nonresponse were evident in a number of the participating countries. As is customary in IEA studies, ICILS 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 ICILS survey populations in accordance with the recommendations of the sampling referee (Jean Dumais, Statistics Canada) 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 bias risks. Because very little information about nonrespondents during ICILS was available, it was not possible in most countries to quantify the risk or bias of estimates due to nonparticipation. To

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 (%)
Australia	90.9	91.3	86.5	78.6	79.0
Chile	95.1	100.0	95.9	91.2	95.9
Croatia	99.6	99.6	96.5	96.0	96.0
Czech Republic	99.3	100.0	99.9	99.2	99.9
Denmark	32.8	58.2	85.5	28.0	49.7
Germany	66.0	81.7	79.5	52.5	64.9
Hong Kong SAR	65.0	70.8	82.2	53.5	58.3
Korea, Republic of	100.0	100.0	99.9	99.9	99.9
Lithuania	91.2	96.8	88.4	80.7	85.6
Netherlands	41.6	64.9	76.3	31.7	49.5
Norway (Grade 9)	70.8	77.6	83.1	58.9	64.5
Poland	86.4	99.4	94.1	81.3	93.6
Russian Federation	99.9	99.9	98.5	98.4	98.4
Slovak Republic	93.1	99.5	98.2	91.4	97.7
Slovenia	88.2	94.8	92.9	82.0	88.1
Switzerland	20.9	36.6	74.2	15.5	27.2
Thailand	79.8	89.0	95.9	76.5	85.4
Turkey	99.1	100.0	95.8	94.9	95.8
Benchmarking participants					
City of Buenos Aires, Argentina	49.5	49.5	77.8	38.6	38.6
Newfoundland and Labrador, Canada	85.8	85.8	92.6	79.4	79.4
Ontario, Canada	73.3	77.4	92.9	68.1	71.9

overcome this, and in addition to the overall participation rate requirements described below, ICILS established strict standards for minimum within-school participation: data from schools with a response rate of less than half (50%) of the 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.

The within-school response rate was computed separately for the student survey and the teacher survey; hence, a school may have counted 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 nonresponse (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 & Byrnes, 2012; Hancock, Shepherd, Lawrence, & Zubrick, 2013). Consequently, the likelihood of biased results may increase as the within-school response rate decreases. A sampled school was regarded as a "participating school" in the student survey if at least 50 percent of its sampled students participated. If the response rate was lower, student data from the affected school were disregarded.

Whenever there was evidence that the survey operation procedures in a school had not been properly followed, that school was also regarded as nonparticipating. For example, if a school failed to list all eligible students for the student sample selection and therefore risked 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 nonresponse (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. As presented later in this chapter, ICILS data suggested that in a number of countries the likelihood of teachers responding to the survey depended on their gender, age, and subject domain. In order to help reduce nonresponse bias, ICILS only regarded a school as a "participating school" in the teacher survey if at least 50 percent of that school's sampled teachers participated. If the response rate was lower, teacher data from the affected school were disregarded.

If a school failed to follow the survey operation procedures properly, ICILS regarded it as nonparticipating. For example, if a school failed to list all eligible teachers for the teacher sample selection, or if it had not followed the standard teacher selection procedures, that school's teacher data were also not included in the final database.

Country-level participation requirements

Three categories were defined for sampling participation:

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

Sampling participation categories for the teacher survey were identical to the ones in the student survey. The results from ICILS 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. Because nonresponse 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. Although there were no participation requirements for reporting school-level data, the participation rate for the school survey was above 85 percent for all countries placed in Category 1 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 2013

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 has either

- An unweighted school response rate with replacement of at least 85 percent (after rounding to the nearest whole percent) and an unweighted overall student/teacher response rate (after rounding) of at least 85 percent

or

- A weighted school response rate with replacement of at least 85 percent (after rounding to 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 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 research team considered it necessary to make readers of the international reports aware of the increased potential for bias, regardless of whether such a bias was actually introduced. In accordance with the sample participation categories, the survey results were reported 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 in the international reports.

During the student survey, six countries failed to meet the sampling participation requirements for Category 1, while eight countries failed to comply with these sampling participation standards during the teacher survey. In Switzerland and the City of Buenos Aires (Argentina), the teacher participation rate was so low that the ICILS joint management committee decided it would be impossible to make inferences from

sample data to population characteristics. These countries were therefore not included in the analyses of teacher data in the international reports.

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

Analysis of nonresponse

As already pointed out earlier in this chapter, nonresponse always holds the potential for biased results. Bias can be substantial when the response rates are low and when the difference in outcome variables between respondents and nonrespondents is large.

Let us illustrate this relationship with an extreme example. Imagine a scenario where all students with a low socioeconomic background refuse to participate in the survey while all others do participate. As presented in Table 4.3 of the ICILS international report (Fraillon et al., 2014), the average CIL scores of students increased with higher parental educational attainment in all countries. Hence, the CIL score for our imaginary country would be overestimated because it would actually represent only students with medium or high socioeconomic backgrounds.

The problem at hand is that no or very little information is known, accessible, or collectable about sampled units that refuse to participate in a survey. Other than characteristics determining stratum membership, ICILS did not collect information from nonparticipating schools or their students and teachers. As such, bias arising

Table 7.5: Achieved participation categories by country

Country	Participation Category	
	Student survey	Teacher survey
Australia	1	1
Chile	1	1
Croatia	1	1
Czech Republic	1	1
Denmark	3	3
Germany	2	3
Hong Kong SAR	3	3
Korea, Republic of	1	1
Lithuania	1	1
Netherlands	3	3
Norway (Grade 9)	1	3
Poland	1	1
Russian Federation	1	1
Slovak Republic	1	1
Slovenia	1	1
Switzerland	3	3 *
Thailand	1	1
Turkey	1	1
Benchmarking participants		
City of Buenos Aires, Argentina	3	3 *
Newfoundland and Labrador, Canada	1	1
Ontario, Canada	1	3

Note: *Due to extremely low participation rates, no results were reported and no data have been published for public use.

from school nonparticipation could not be estimated. However, some information was collected from all eligible individuals within participating schools before the within-school sampling, thus allowing for some limited analysis of differential student response within participating schools.

For students, gender information was generally collected prior to student sampling. Table 7.6 presents the numbers and percentages of responding and nonresponding sampled students by gender. Given that gender had a sizable effect on ICILS' key main outcome variables, it is reassuring to see that a dependency between response likelihood and gender was only detectable in two countries—Denmark and Turkey.[3] In both countries, the proportion of boys responding to the survey was higher than the proportion of girls doing so. However, in Denmark, a post-stratification weight adjustment by gender would have changed the total CIL average score by as little as 0.5 score points. Therefore, the idea of building in a respective correction factor was abandoned in favor of applying the same parsimonious method for all countries. In Turkey, the change in total average CIL score points would have been almost undetectable because the data showed no statistical differences in terms of CIL scale scores between boys and girls.

ICILS asked schools to provide information on gender, age, and the main subject domain of teaching for all teachers before the within-school teacher selection. Tables 7.7 to 7.9 give numbers and percentages of responding and nonresponding sampled teachers by these characteristics.

As shown in Table 7.7, six countries showed significantly different response patterns for gender groups. They were Australia, Croatia, Lithuania, the Netherlands, Norway (Grade 9), and Slovenia. In all these countries, the participation rate among female teachers was higher than among male teachers. The biggest difference was recorded in Lithuania, where the response rate of male teachers was 8.3 percentage points lower than the response rate of female teachers. Researchers performing secondary analysis of the ICILS data therefore need to monitor gender patterns in these countries and to interpret results with some caution.

Eight countries showed an association between response and age groups (see Table 7.8), with older teachers being more likely than younger teachers not to respond to the survey. However, five of these eight countries had generally very high within-school response rates (> 90%), so even if the data showed some dependency between willingness to participate and age group, the bias could be considered low or even negligible. This pattern applied to Croatia, the Czech Republic, the Russian Federation, the Slovak Republic, and Slovenia.

For two other countries out of the eight, Australia and Ontario (Canada), age information was missing from the teacher-listing forms for a high percentage of teachers, making such analysis unreliable. Germany showed quite clear age-related response patterns, especially in terms of teachers in the higher age groups being less likely than those in the lower groups to participate in the survey. Given that age played a major role in a number of teacher-related scales presented in the report, caution should be exercised when reporting and interpreting results across age groups.

3 A Chi² test was employed to detect differences in response patterns between groups.

Table 7.6: Nonresponse by gender, student survey

Country	Gender	Number of Participating Students*	Number of Nonparticipating Students	Participating Students (%)	Nonparticipating Students (%)	CHI Squared Significant
Australia	Missing	0	1	0.0	100.0	
	Girls	2705	344	88.7	11.3	no
	Boys	2664	332	88.9	11.1	
Chile	Girls	1552	100	93.9	6.1	no
	Boys	1628	76	95.5	4.5	
Croatia	Girls	1425	311	82.1	17.9	no
	Boys	1488	279	84.2	15.8	
Czech Republic	Girls	1556	82	95.0	5.0	no
	Boys	1510	85	94.7	5.3	
Denmark	Girls	902	138	86.7	13.3	yes
	Boys	950	97	90.7	9.3	
Germany	Missing	0	17	0.0	100.0	
	Girls	1122	256	81.4	18.6	no
	Boys	1148	264	81.3	18.7	
Hong Kong SAR	Missing	1	25	3.8	96.2	
	Girls	1023	147	87.4	12.6	no
	Boys	1120	133	89.4	10.6	
Korea, Republic of	Girls	1408	20	98.6	1.4	no
	Boys	1480	23	98.5	1.5	
Lithuania	Girls	1346	86	94.0	6.0	no
	Boys	1414	89	94.1	5.9	
Netherlands	Missing	1	71	1.4	98.6	
	Girls	1053	112	90.4	9.6	no
	Boys	1151	108	91.4	8.6	
Norway (Grade 9)	Girls	1222	120	91.1	8.9	no
	Boys	1214	104	92.1	7.9	
Poland	Missing	0	1	0.0	100.0	
	Girls	1371	232	85.5	14.5	no
	Boys	1501	259	85.3	14.7	
Russian Federation	Girls	1839	47	97.5	2.5	no
	Boys	1807	56	97.0	3.0	
Slovak Republic	Girls	1482	93	94.1	5.9	no
	Boys	1523	84	94.8	5.2	
Slovenia	Girls	1818	177	91.1	8.9	no
	Boys	1932	149	92.8	7.2	
Switzerland	Missing	0	3	0.0	100.0	
	Girls	1567	118	93.0	7.0	no
	Boys	1677	125	93.1	6.9	
Thailand	Girls	1814	25	98.6	1.4	no
	Boys	1832	64	96.6	3.4	
Turkey	Girls	1294	64	95.3	4.7	yes
	Boys	1402	106	93.0	7.0	
Benchmarking participants						
City of Buenos Aires, Argentina	Missing	3	2	60.0	40.0	
	Girls	730	130	84.9	15.1	no
	Boys	680	122	84.8	15.2	
Newfoundland and Labrador, Canada	Missing	0	1	0.0	100.0	
	Girls	826	103	88.9	11.1	no
	Boys	733	94	88.6	11.4	
Ontario, Canada	Girls	1689	133	92.7	7.3	no
	Boys	1707	134	92.7	7.3	

Note: *Some of the students counted here were later treated as nonparticipants because they belonged to schools with low response rates (< 50%).

Table 7.7: Nonresponse by gender, teacher survey

Country	Gender	Number of Participating Teachers*	Number of Nonparticipating Teachers	Participating Teachers (%)	Nonparticipating Teachers (%)	CHI Squared Significant
Australia	Missing	0	2	0.0	100.0	
	Women	2252	409	84.6	15.4	yes
	Men	1382	340	80.3	19.7	
Chile	Women	1094	55	95.2	4.8	no
	Men	706	20	97.2	2.8	
Croatia	Women	1925	66	96.7	3.3	yes
	Men	656	37	94.7	5.3	
Czech Republic	Women	1577	2	99.9	0.1	no
	Men	549	1	99.8	0.2	
Denmark	Women	498	232	68.2	31.8	no
	Men	343	153	69.2	30.8	
Germany	Women	898	342	72.4	27.6	no
	Men	601	224	72.8	27.2	
Hong Kong SAR	Missing	0	70	0.0	100.0	
	Women	833	240	77.6	22.4	no
	Men	563	159	78.0	22.0	
Korea, Republic of	Women	1528	1	99.9	0.1	no
	Men	661	1	99.8	0.2	
Lithuania	Women	1846	197	90.4	9.6	yes
	Men	325	71	82.1	17.9	
Netherlands	Missing	2	264	0.8	99.2	
	Women	601	171	77.8	22.2	yes
	Men	559	197	73.9	26.1	
Norway (Grade 9)	Women	771	239	76.3	23.7	yes
	Men	445	167	72.7	27.3	
Poland	Women	1659	92	94.7	5.3	no
	Men	569	47	92.4	7.6	
Russian Federation	Women	2334	50	97.9	2.1	no
	Men	396	11	97.3	2.7	
Slovak Republic	Women	1677	30	98.2	1.8	no
	Men	468	18	96.3	3.7	
Slovenia	Women	2182	181	92.3	7.7	yes
	Men	624	64	90.7	9.3	
Thailand	Women	1357	112	92.4	7.6	no
	Men	770	90	89.5	10.5	
Turkey	Women	987	38	96.3	3.7	no
	Men	900	40	95.7	4.3	
Benchmarking participants						
City of Buenos Aires, Argentina	Missing	0	2	0.0	100.0	
	Women	579	341	62.9	37.1	no
	Men	217	141	60.6	39.4	
Newfoundland and Labrador, Canada	Women	227	46	83.2	16.8	no
	Men	187	41	82.0	18.0	
Ontario, Canada	Missing	0	11	0.0	100.0	
	Women	277	69	80.1	19.9	no
	Men	186	55	77.2	22.8	

Note: *Some of the teachers counted here were later treated as nonparticipants because they belonged to schools with low response rates (< 50%).

Table 7.8: Nonresponse by age group, teacher survey

Country	Age	Number of Participating Teachers*	Number of Nonparticipating Teachers	Participating Teachers (%)	Nonparticipating Teachers (%)	Chi Squared Significant
Australia	1) Under 25	100	20	83.3	16.7	yes
	2) 25 to 29 years	416	70	85.6	14.4	
	3) 30 to 39 years	717	109	86.8	13.2	
	4) 40 to 49 years	672	109	86.0	14.0	
	5) 50 to 59 years	599	105	85.1	14.9	
	6) 60 or over	215	49	81.4	18.6	
	Missing age	915	289	76.0	24.0	
Chile	1) Under 25	20	3	87.0	13.0	no
	2) 25 to 29 years	305	10	96.8	3.2	
	3) 30 to 39 years	515	29	94.7	5.3	
	4) 40 to 49 years	428	10	97.7	2.3	
	5) 50 to 59 years	398	15	96.4	3.6	
	6) 60 or over	134	8	94.4	5.6	
Croatia	1) Under 25	4	0	100.0	0.0	yes
	2) 25 to 29 years	258	2	99.2	0.8	
	3) 30 to 39 years	817	23	97.3	2.7	
	4) 40 to 49 years	630	24	96.3	3.7	
	5) 50 to 59 years	480	24	95.2	4.8	
	6) 60 or over	354	30	92.2	7.8	
	Missing age	38	0	100.0	0.0	
Czech Republic	1) Under 25	3	0	100.0	0.0	yes
	2) 25 to 29 years	159	0	100.0	0.0	
	3) 30 to 39 years	575	3	99.5	0.5	
	4) 40 to 49 years	620	0	100.0	0.0	
	5) 50 to 59 years	557	0	100.0	0.0	
	6) 60 or over	157	0	100.0	0.0	
	Missing age	55	0	100.0	0.0	
Denmark	1) Under 25	1	0	100.0	0.0	no
	2) 25 to 29 years	36	14	72.0	28.0	
	3) 30 to 39 years	278	127	68.6	31.4	
	4) 40 to 49 years	206	106	66.0	34.0	
	5) 50 to 59 years	197	87	69.4	30.6	
	6) 60 or over	66	44	60.0	40.0	
	Missing age	57	7	89.1	10.9	
Germany	1) Under 25	2	0	100.0	0.0	yes
	2) 25 to 29 years	110	32	77.5	22.5	
	3) 30 to 39 years	347	121	74.1	25.9	
	4) 40 to 49 years	384	130	74.7	25.3	
	5) 50 to 59 years	427	179	70.5	29.5	
	6) 60 or over	195	104	65.2	34.8	
	Missing age	34	0	100.0	0.0	
Hong Kong SAR	1) Under 25	7	4	63.6	36.4	no
	2) 25 to 29 years	58	12	82.9	17.1	
	3) 30 to 39 years	109	30	78.4	21.6	
	4) 40 to 49 years	118	36	76.6	23.4	
	5) 50 to 59 years	52	18	74.3	25.7	
	6) 60 or over	5	0	100.0	0.0	
	Missing age	1047	369	73.9	26.1	
Korea, Republic of	1) Under 25	11	0	100.0	0.0	no
	2) 25 to 29 years	208	0	100.0	0.0	
	3) 30 to 39 years	617	0	100.0	0.0	
	4) 40 to 49 years	727	1	99.9	0.1	

Table 7.8: Nonresponse by age group, teacher survey (contd.)

Country	Age	Number of Participating Teachers*	Number of Nonparticipating Teachers	Participating Teachers (%)	Nonparticipating Teachers (%)	Chi Squared Significant
Korea, Republic of	5) 50 to 59 years	585	0	100.0	0.0	
	6) 60 or over	20	0	100.0	0.0	
	Missing age	21	1	95.5	4.5	
Lithuania	1) Under 25	8	3	72.7	27.3	
	2) 25 to 29 years	107	11	90.7	9.3	
	3) 30 to 39 years	391	49	88.9	11.1	
	4) 40 to 49 years	640	63	91.0	9.0	no
	5) 50 to 59 years	761	98	88.6	11.4	
	6) 60 or over	224	33	87.2	12.8	
	Missing age	40	11	78.4	21.6	
Netherlands	1) Under 25	30	7	81.1	18.9	
	2) 25 to 29 years	85	39	68.5	31.5	
	3) 30 to 39 years	156	67	70.0	30.0	
	4) 40 to 49 years	127	50	71.8	28.2	no
	5) 50 to 59 years	187	69	73.0	27.0	
	6) 60 or over	58	39	59.8	40.2	
	Missing age	519	361	59.0	41.0	
Norway (Grade 9)	1) Under 25	11	1	91.7	8.3	
	2) 25 to 29 years	93	31	75.0	25.0	
	3) 30 to 39 years	347	90	79.4	20.6	
	4) 40 to 49 years	352	114	75.5	24.5	no
	5) 50 to 59 years	233	85	73.3	26.7	
	6) 60 or over	174	73	70.4	29.6	
	Missing age	6	12	33.3	66.7	
Poland	1) Under 25	4	1	80.0	20.0	
	2) 25 to 29 years	116	11	91.3	8.7	
	3) 30 to 39 years	649	29	95.7	4.3	
	4) 40 to 49 years	654	42	94.0	6.0	no
	5) 50 to 59 years	459	30	93.9	6.1	
	6) 60 or over	34	2	94.4	5.6	
	Missing age	312	24	92.9	7.1	
Russian Federation	1) Under 25	69	0	100.0	0.0	
	2) 25 to 29 years	165	7	95.9	4.1	
	3) 30 to 39 years	439	5	98.9	1.1	
	4) 40 to 49 years	823	17	98.0	2.0	yes
	5) 50 to 59 years	813	20	97.6	2.4	
	6) 60 or over	287	12	96.0	4.0	
	Missing age	134	0	100.0	0.0	
Slovak Republic	1) Under 25	4	0	100.0	0.0	
	2) 25 to 29 years	231	5	97.9	2.1	
	3) 30 to 39 years	657	9	98.6	1.4	
	4) 40 to 49 years	518	14	97.4	2.6	yes
	5) 50 to 59 years	554	8	98.6	1.4	
	6) 60 or over	178	12	93.7	6.3	
	Missing age	3	0	100.0	0.0	
Slovenia	1) Under 25	0	0	0.0	0.0	
	2) 25 to 29 years	130	8	94.2	5.8	
	3) 30 to 39 years	875	59	93.7	6.3	
	4) 40 to 49 years	793	55	93.5	6.5	yes
	5) 50 to 59 years	910	91	90.9	9.1	
	6) 60 or over	52	9	85.2	14.8	
	Missing age	46	23	66.7	33.3	

Table 7.8: Nonresponse by age group, teacher survey (contd.)

Country	Age	Number of Participating Teachers*	Number of Nonparticipating Teachers	Participating Teachers (%)	Nonparticipating Teachers (%)	Chi Squared Significant
Thailand	1) Under 25	63	1	98.4	1.6	
	2) 25 to 29 years	334	23	93.6	6.4	
	3) 30 to 39 years	641	55	92.1	7.9	
	4) 40 to 49 years	420	35	92.3	7.7	no
	5) 50 to 59 years	389	42	90.3	9.7	
	6) 60 or over	37	1	97.4	2.6	
	Missing age	243	45	84.4	15.6	
Turkey	1) Under 25	82	5	94.3	5.7	
	2) 25 to 29 years	513	14	97.3	2.7	
	3) 30 to 39 years	851	42	95.3	4.7	
	4) 40 to 49 years	299	12	96.1	3.9	no
	5) 50 to 59 years	106	4	96.4	3.6	
	6) 60 or over	8	1	88.9	11.1	
	Missing age	28	0	100.0	0.0	
Benchmarking participants						
City of Buenos Aires, Argentina	1) Under 25	9	2	81.8	18.2	
	2) 25 to 29 years	54	30	64.3	35.7	
	3) 30 to 39 years	160	103	60.8	39.2	
	4) 40 to 49 years	226	118	65.7	34.3	no
	5) 50 to 59 years	210	140	60.0	40.0	
	6) 60 or over	26	23	53.1	46.9	
	Missing age	111	156	41.6	58.4	
Newfoundland and Labrador, Canada	1) Under 25	2	2	50.0	50.0	
	2) 25 to 29 years	42	9	82.4	17.6	
	3) 30 to 39 years	71	17	80.7	19.3	
	4) 40 to 49 years	88	19	82.2	17.8	no
	5) 50 to 59 years	43	8	84.3	15.7	
	6) 60 or over	4	0	100.0	0.0	
	Missing age	164	32	83.7	16.3	
Ontario, Canada	1) Under 25	2	0	100.0	0.0	
	2) 25 to 29 years	24	3	88.9	11.1	
	3) 30 to 39 years	90	17	84.1	15.9	
	4) 40 to 49 years	64	17	79.0	21.0	yes
	5) 50 to 59 years	31	6	83.8	16.2	
	6) 60 or over	3	1	75.0	25.0	
	Missing age	249	91	73.2	26.8	

Note: *Some of the teachers counted here were later treated as nonparticipants because they belonged to schools with low response rates (< 50%).

Finally, Table 7.9 displays the response patterns of teachers by their main subject domains. Significant differences were detected for Croatia, Germany, Lithuania, Poland, and Slovenia, but no general patterns could be discerned across countries. Again, little concern with respect to bias needs to be raised for countries with high general response rates (Croatia, Lithuania, Poland, and Slovenia). However, German teacher data, in particular, should be, as already noted above, analyzed and interpreted with caution.

Table 7.9: Nonresponse by main subject domain, teacher survey

Country	Main Subject Domain	Number of Participating Teachers*	Number of Nonparticipating Teachers	Participating Teachers (%)	Nonparticipating Teachers (%)	Chi Squared Significant
Australia	Missing	79	43	64.8	35.2	
	1	897	165	84.5	15.5	
	2	405	86	82.5	17.5	no
	3	954	183	83.9	16.1	
	4	1299	274	82.6	17.4	
Chile	1	381	16	96.0	4.0	
	2	183	10	94.8	5.2	no
	3	445	20	95.7	4.3	
	4	791	29	96.5	3.5	
Croatia	Missing	1	0	100.0	0.0	
	1	911	32	96.6	3.4	
	2	352	11	97.0	3.0	yes
	3	695	24	96.7	3.3	
	4	622	36	94.5	5.5	
Czech Republic	1	716	2	99.7	0.3	
	2	311	0	100.0	0.0	no
	3	630	1	99.8	0.2	
	4	469	0	100.0	0.0	
Denmark	Missing	19	0	100.0	0.0	
	1	393	181	68.5	31.5	
	2	91	42	68.4	31.6	no
	3	271	128	67.9	32.1	
	4	67	34	66.3	33.7	
Germany	1	544	209	72.2	27.8	
	2	155	70	68.9	31.1	yes
	3	474	145	76.6	23.4	
	4	326	142	69.7	30.3	
Hong Kong SAR	Missing	82	71	53.6	46.4	
	1	471	155	75.2	24.8	
	2	222	64	77.6	22.4	no
	3	312	87	78.2	21.8	
	4	309	92	77.1	22.9	
Korea, Republic of	Missing	1	0	100.0	0.0	
	1	632	0	100.0	0.0	
	2	347	0	100.0	0.0	no
	3	560	2	99.6	0.4	
	4	649	0	100.0	0.0	
Lithuania	1	885	105	89.4	10.6	
	2	243	32	88.4	11.6	yes
	3	490	53	90.2	9.8	
	4	553	78	87.6	12.4	
Netherlands	Missing	481	352	57.7	42.3	
	1	221	97	69.5	30.5	
	2	122	40	75.3	24.7	no
	3	151	65	69.9	30.1	
	4	187	78	70.6	29.4	
Norway (Grade 9)	Missing	2	1	66.7	33.3	
	1	492	160	75.5	24.5	
	2	101	37	73.2	26.8	no
	3	302	78	79.5	20.5	
	4	319	130	71.0	29.0	

Table 7.9: Nonresponse by main subject domain, teacher survey (contd.)

Country	Main Subject Domain	Number of Participating Teachers*	Number of Nonparticipating Teachers	Participating Teachers (%)	Nonparticipating Teachers (%)	Chi Squared Significant
Poland	Missing	11	0	100.0	0.0	
	1	662	26	96.2	3.8	
	2	326	25	92.9	7.1	yes
	3	598	35	94.5	5.5	
	4	631	53	92.3	7.7	
Russian Federation	1	690	14	98.0	2.0	
	2	325	11	96.7	3.3	no
	3	850	16	98.2	1.8	
	4	865	20	97.7	2.3	
Slovak Republic	Missing	3	0	100.0	0.0	
	1	714	12	98.3	1.7	
	2	338	9	97.4	2.6	no
	3	610	12	98.1	1.9	
	4	480	15	97.0	3.0	
Slovenia	1	874	69	92.7	7.3	
	2	356	22	94.2	5.8	yes
	3	811	54	93.8	6.2	
	4	765	100	88.4	11.6	
Thailand	Missing	103	10	91.2	8.8	
	1	493	54	90.1	9.9	
	2	242	20	92.4	7.6	no
	3	499	46	91.6	8.4	
	4	790	72	91.6	8.4	
Turkey	Missing	33	0	100.0	0.0	
	1	499	19	96.3	3.7	
	2	208	11	95.0	5.0	no
	3	406	15	96.4	3.6	
	4	741	33	95.7	4.3	
Benchmarking participants						
City of Buenos Aires, Argentina	Missing	81	96	45.8	54.2	
	1	105	74	58.7	41.3	
	2	152	97	61.0	39.0	no
	3	145	100	59.2	40.8	
	4	313	205	60.4	39.6	
Newfoundland and Labrador, Canada	Missing	1	0	100.0	0.0	
	1	113	23	83.1	16.9	
	2	44	7	86.3	13.7	no
	3	120	24	83.3	16.7	
	4	136	33	80.5	19.5	
Ontario, Canada	Missing	29	18	61.7	38.3	
	1	202	53	79.2	20.8	
	2	18	6	75.0	25.0	no
	3	114	28	80.3	19.7	
	4	100	30	76.9	23.1	

Key:
Main subject domains
1: Language Arts
2: Human Sciences
3: Mathematics and Sciences
4: Other
Note: *Some of the teachers counted here were treated as nonparticipants later on because they belonged to schools with low response rates (< 50%).

Summary

The several sets of weights computed for ICILS data reflect not only the varying selection probabilities for the selected students and teachers but also the varying patterns of nonparticipation between strata and within schools. All findings presented in ICILS reports are based on weighted data. Any secondary analysis should only be undertaken using weighted data in order to obtain accurate population estimates.

As student or teacher response rates within a school drop, the likelihood of bias increases. Therefore, ICILS defined the minimum student and teacher response rate requirements within each school that would guarantee their inclusion in the database. ICILS also requested that minimum participation rates at school level and overall in a country be determined and that these determinations should govern whether or not data for these countries were included in the reporting tables.

ICILS calculated unweighted and weighted participation rates for the student and the teacher surveys. Results pertaining to countries that did not meet the IEA requirements were annotated or reported in separate sections of the tables in the ICILS international report (Fraillon et al., 2014). Very little information was available about nonresponding students and teachers, and analysis pertaining to nonresponse confirmed that results from countries with relatively low participation rates need to be interpreted with caution.

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.

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.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Franklin, S., & Walker C. (Eds.). (2003). *Survey methods and practices.* Ottawa, ON, Canada: Statistics Canada, Social Survey Methods Division.

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.* Report prepared for the Department of Education, Employment and Workplace Relations, Canberra, ACT, Australia.

IEA Data Processing and Research Center. (2012). Windows® Within-School Sampling Software (WinW3S) [Computer software]. Hamburg, Germany: Author.

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

Meinck, S. (2015, February). Computing sampling weights in large-scale assessments in education. *Survey methods: Insights from the field (online journal*). Retrieved from http://surveyinsights.org/

Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 118–151). New York, NY: CRC Press.

CHAPTER 8:

ICILS Field Operations

Michael Jung and Ralph Carstens

Introduction

Successful administration of the ICILS 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 crossnational 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 student instruments on computer, which was a first in IEA studies.

The ICILS international study center (ISC) at the Australian Council for Educational Research (ACER) in cooperation with the IEA Secretariat and IEA Data Processing and Research Center (IEA DPC) 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 procedures developed for previous 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, 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 they 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 had to take when establishing the study in their country was to appoint a national research coordinator (NRC). The NRC acted as the main contact person for all those involved in ICILS within the country. He or she also represented the country 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, 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 school coordinators from their own members of staff, for example. 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.

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 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 for 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 included the following:

- *Unit 1: Sampling Schools* (IEA, 2012a): This manual specified the actions and procedures required to develop a national sampling plan in compliance with the international ICILS sample design.
- *Unit 2: Working with Schools* (IEA, 2012b): This contained information about how to work with schools in order to plan for successful administration of the ICILS instruments.
- *Unit 3: Instrument Preparation* (IEA, 2012c): This unit described the processes involved in preparing the ICILS instruments for production and use in countries.
- *Unit 4: Data Collection and Quality Monitoring Procedures* (IEA, 2012d): This document dealt with the processes involved in preparing for, supporting, and monitoring the ICILS data collection in schools.

- *Unit 5: Post Collection Data Capture, Data Upload, Scoring, and Parental Occupation Coding* (IEA, 2012e): This unit 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.
- *School Coordinator Manual* (IEA, 2012f), subject to translation: This manual described the role and responsibilities of the school coordinator, the main contact person within each participating school.
- *Test Administrator Manual* (IEA, 2012g), subject to translation: This manual described the role and responsibilities of the test administrator, whose work included administration of the student assessment.
- *National Quality Control Monitor Manual* (IEA, 2012h): This provided national quality control monitors (NQCMs) with information about ICILS, 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.
- *International Quality Control Monitor Manual* (IEA, 2013): This provided international quality control monitors (IQCMs) with information about ICILS, their role and responsibilities during the project, and the timelines, actions, and procedures to be followed in order to carry out the international quality control programs.
- *Scoring Guides for Constructed-Response Items* (IEA, 2012i), subject to translation: These provided detailed and explicit guidelines on how to score each constructed-response item.
- *Student Instrument Translation Software Manual* (IEA, 2012j): This described the use of the translation software platform that allowed for the adaptation, translation, and verification of the ICILS student instruments in their computer-based format.
- *Compatibility Test and School Computer Resources Survey: Instructions for NRCs* (IEA, 2012k) and *Compatibility Test and School Computer Resources Survey: Instructions for School Coordinators* (IEA, 2012l): These documents addressed whether computers in the sampled schools could be used for the ICILS 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:

- *The ICILS Translation Software:* This web-based application supported translation, translation verification, and layout verification of the student instruments (test modules, tutorial, and questionnaire). This software also allowed cultural adaptations and verification for English-speaking countries that did not require translations.
- *The ICILS Student Test Software:* This software was used to administer the computer-based ICILS 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 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.

- *The ICILS Scoring Software:* 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, scoring training responses, scoring student responses, and flagging responses for review scoring.
- *The IEA Windows® Within-School Sampling Software* (IEA WinW3S): This enabled the ICILS 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.
- *The IEA SurveySystem* (IEA OSS): This software enabled text passages on the paper questionnaires to be transferred to online questionnaires, while taking national adaptations to be made to the questionnaires into account. The software also made it possible to deliver these online versions to respondents.
- *The IEA Data Management Expert* (IEA DME): This software facilitated the entering of paper questionnaire data. The DME software also allowed for national adaptations to be made to the questionnaires.

In addition to its work preparing the software and manuals, the IEA DPC conducted a data-management seminar designed to train national center staff on all procedures and the software supporting these, namely IEA WinW3S, IEA SurveySystem, IEA Data Management Expert, and the ICILS Student Test Software. This seminar was combined with training in scoring, during which national center staff received instruction on how to use the ICILS Scoring Software. Instructions for using the ICILS Translation Software 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 identification 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 teacher identification 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, 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 (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.

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. 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 test, 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.

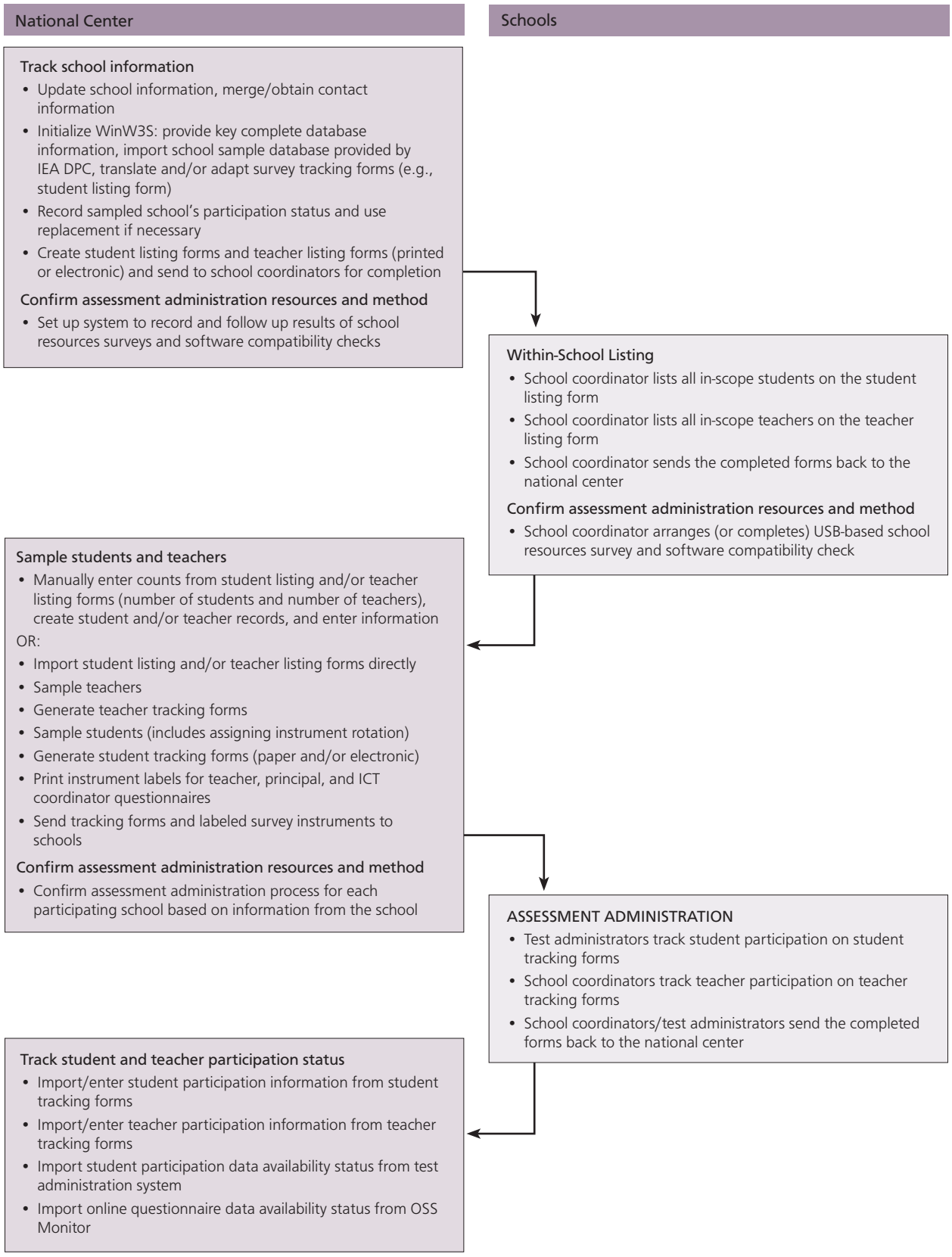
National centers sent a *student listing form* to each school coordinator and asked him or her to provide information on all eligible target-grade students in the school. School coordinators collected details about these students, such as their names (if country regulations allowed national centers to be given names), 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 him or her to provide information on all eligible target-grade teachers within the school.

1 Although all students enrolled in the target grade were part of the target population, ICILS recognized that some student exclusions were necessary because of physical or intellectual disability or in cases where there were nonnative language speakers not proficient enough 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 flag 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 for working with schools



The school coordinators listed the eligible target-grade teachers and provided details about these teachers, such as their names (if country regulations allowed for names to be sent to the national center), birth month and year, and gender. The national centers used the collected information to sample teachers within the schools.

Preparing the computer-based test delivery at schools

Because ICILS 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 test 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 for the USB-based student test to run correctly (e.g., altering the configuration of computers or using a laptop local server connected to the school local area network).

The process of administering the compatibility test 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 test file, on a USB stick to school coordinators.

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

- Run the USB compatibility test on every computer that was to be used for the ICILS 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 in number 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 centers sent teacher questionnaires to each teacher listed on the *teacher tracking form*, in each school. They also sent a school 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 also 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 testing, 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 Student Test Software. Administration of the ICILS assessment consisted of three parts, the first two of which required students to complete the first and second student test modules and the third to answer the student questionnaire. Test administrators were requested to document student participation on the student tracking forms.

During 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. 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 at least 20 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*. Table 8.2 details the time allotted to the different parts of the student assessment.

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

Table 8.2: Timing of the ICILS assessment

Activities	Length
Preparation of students, reading of instructions, and administering the tutorial	20 minutes (approximate)
Administering the student assessment—first module	30 minutes (exact)
Short break	5 minutes (max.)
Administering the student assessment—second module	30 minutes (exact)
Short break	5 minutes (max.)
Administering the student questionnaire	Approx. 20 minutes
Collecting the assessment materials and ending the session	Approx. 5 minutes
Total	Approx. 2 hours

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

ICILS 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 field trial were allowed to use the online option during the main survey. All countries (with the exception of the City of Buenos Aires) used the online administration mode for at least some of 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 software as described in more detail in Chapter 5 of this report.

The electronic versions of the ICILS 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 the paper questionnaire as a default to respondents. This decision was based on the centers' and the 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 not being willing 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 to resume answering the questionnaire at the question they had last responded to in their previous session. Answers were automatically saved whenever respondents moved to another question, and respondents could change any answer at any time before completing the questionnaire. During the administration, the national center was available for support; the center, in turn, could contact the IEA DPC if it was 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 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 the IEA DPC, there was no need for separate data entry.

Table 8.3 shows the (weighted) percentages of principal, ICT coordinator, and teacher questionnaires that were completed online.

Table 8.3: Percentage of questionnaires administered online

Country	Principal Questionnaire (%)	ICT Coordinator Questionnaire (%)	Teacher Questionnaire (%)
Australia	100.0	100.0	100.0
Chile	87.1	90.2	89.4
Croatia	9.6	9.6	10.9
Czech Republic	100.0	100.0	100.0
Denmark	100.0	100.0	100.0
Germany	49.6	54.8	34.7
Hong Kong SAR	100.0	100.0	100.0
Korea, Republic of	100.0	100.0	100.0
Lithuania	100.0	100.0	100.0
Netherlands	47.1	100.0	29.7
Norway	100.0	100.0	100.0
Poland	19.5	27.7	23.8
Russian Federation	94.8	94.2	95.7
Slovak Republic	100.0	100.0	100.0
Slovenia	100.0	100.0	99.9
Switzerland	100.0	100.0	–
Thailand	89.4	89.9	86.9
Turkey	100.0	100.0	100.0
Benchmarking participants			
City of Buenos Aires, Argentina			–
Newfoundland and Labrador, Canada	100.0	100.0	100.0
Ontario, Canada	98.9	98.9	99.2

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 SurveySystem. Because the survey activities questionnaire, unlike the other ICILS 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 strength and weaknesses of the ICILS assessment materials (e.g., test instruments, manuals, scoring guides, and software) as well as countries' experiences with the ICILS 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 for use in future ICILS cycles.

The IEA DPC sent the NRCs individual login information and internet links for accessing the online questionnaires. Before submitting the responses to the IEA DPC, NRCs could go back and change their answers if necessary.

Scoring the assessment and checking scorer reliability

Scoring the assessment

Nine of the ICILS assessment items were constructed-response items. The four large tasks were scored against a total of 35 criteria. One of the large-task criteria was automatically scored by the ICILS delivery system, which provided suggested scores for a further three criteria that human scorers could then confirm or override. Of the 70 ICILS tasks, 43 were scored by human scorers, and it was critical to the quality of the ICILS results that these tasks were scored in a reliable manner. Reliability was accomplished by providing national centers with explicit scoring guides, extensive training of scoring staff, and continuous monitoring of the quality of the work during scoring procedures.

During the scoring training, which was conducted at the international level, national center staff members learned how to score the constructed-response items and to use the scoring criteria for the large-task items in the ICILS 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 main survey scorer training employed a sample set of student responses collected during the field trial in English-speaking ICILS 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 and familiar with education and who, to the greatest extent possible, had a background in information and computer literacy. 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 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, that a random selection of 20 percent of responses by each scorer 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 to 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

In almost all participating countries, the international field trial was conducted from February 2012 to April 2012. The field trial was crucial to the development of the ICILS assessment instruments and also served the purpose of testing the ICILS survey operations procedures in order to avoid any possible problems during the ICILS 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.

Summary

Considerable effort was made to ensure high standards of quality in the survey procedures for the ICILS 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 the following: sampling and tracking students and teachers within schools, the computer-based student assessment, data capture, and the optional online administration of contextual questionnaires. The international ICILS field trial in 2012 was crucial for testing survey operations procedures in participating countries and contributed to the successful implementation of the main data collection.

References

IEA (2012a). *IEA International Computer and Information Literacy Study (ICILS) 2013. Survey operations procedures, Unit 1, main survey: Sampling schools*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012b). *IEA International Computer and Information Literacy Study (ICILS) 2013. Survey operations procedures, Unit 2, main survey: Working with schools*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012c). *IEA International Computer and Information Literacy Study (ICILS) 2013. Survey operations procedures, Unit 3, main survey: Instrument preparation*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012d). *IEA International Computer and Information Literacy Study (ICILS) 2013. Survey operations procedures, Unit 4, main survey: Data collection and quality monitoring procedures*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012e). *IEA International Computer and Information Literacy Study (ICILS) 2013. Survey operations procedures, Unit 5, main survey: Post collection data capture, data upload, scoring and parental occupation coding*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012f). *IEA International Computer and Information Literacy Study (ICILS) 2013: School coordinator manual*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012g). *IEA International Computer and Information Literacy Study (ICILS) 2012: Test administrator manual*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012h). *IEA International Computer and Information Literacy Study (ICILS) 2013: National quality control monitor manuals*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012i). *IEA International Computer and Information Literacy Study (ICILS) 2013: Scoring guides for constructed-response items*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012j). *IEA International Computer and Information Literacy Study (ICILS) 2013. Student instrument translation software manual*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012k). *IEA International Computer and Information Literacy Study (ICILS) 2013. Compatibility test and school resources survey: Instructions for NRCs*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2012l). *IEA International Computer and Information Literacy Study (ICILS) 2013: Compatibility test and school resources survey: Instructions for school coordinators*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

IEA (2013). *IEA International Computer and Information Literacy Study (ICILS) 2013: International quality control monitor manual*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

CHAPTER 9:

Quality Assurance of the ICILS Data Collection

Paulína Koršňáková and David Ebbs

Introduction

The student assessment of computer and information literacy (CIL) was an authentic, computer-based test delivered on USB sticks or, in some cases, through a laptop server computer connected to the school network. The test was accompanied by a student questionnaire. The ICILS teacher and school questionnaires, administered mainly online, collected information about computer use, computing resources, and relevant policies and practices in the school context.

Considerable effort was made to develop and standardize materials and procedures so that the data collected in each country for ICILS would be comparable across countries to the greatest extent possible. In order to further ensure the quality of the ICILS data, quality assurance became an integral part of the study work and encompassed all major activities generated from the CIL framework. These activities included instrument development, sampling, translation, verification of the national versions of all instruments, field operations, data collection, scaling, analysis, and reporting.

This chapter focuses on the results of two distinct actions of the quality assurance process: the survey activities questionnaire (SAQ), and the international quality control monitoring (IQCM). The SAQ was completed by the national research coordinator (NRC) in each country, with assistance from the national data manager (NDM) and other local staff where necessary. The IQCM was conducted by independent experts (IQC monitors) appointed and trained by the IEA Secretariat. IQC monitors visited 15 of the participating schools in each country on the day of their testing to discuss the ICILS test administration and to observe the ICILS testing sessions.

Survey activities questionnaire

The purpose of the SAQ was to gather information about the implementation of the ICILS 2013 main survey procedures in all participating education systems. This questionnaire collected NRCs' feedback on the strengths and weaknesses of the approaches and materials used in the ICILS 2013 main data collection. All participating education systems except the city of Buenos Aires provided this information. In the case of the two Canadian provinces, one response covered both.

Data from the SAQ constitute important evidence for assessing the quality of the data collection from ICILS 2013 and for improving future cycles of ICILS and other IEA studies.

The SAQ questions addressed the following areas of interest:

- Sampling of schools;
- Contacting schools and recruiting school coordinators;
- Adapting and translating the ICILS assessment materials;
- Assembling and preparing the ICILS materials for administration;

- Preparing online questionnaires;
- Administering the student instruments and contextual questionnaires;
- Administering the online questionnaires;
- Monitoring the quality of the national data collection;
- Scoring open-ended response items;
- Entering and coding occupation data;
- Entering data manually and submitting data;
- Determining the time required for survey activities; and
- Other experiences.

This section reports on those aspects that the independent IQC monitors observed.

The first part of the SAQ collected information on the sampling-related activities. Table 9.1 provides a summary of the responses for this section of the questionnaire.

Table 9.1: Survey activities questionnaire responses: sampling

Question	Yes	No
Were there any conditions or organizational constraints that required deviations from the standard ICILS 2013 main survey within-school sampling design?	5	14
Did you let staff of the sampled schools complete the listing and tracking forms for teachers and students provided by the Within-School Sampling Software (WinW3S)?	15	4
Did you complete student/teacher listing and tracking forms electronically or on paper?		
• *All electronically*	8	0
• *More than half electronically*	8	0
• *More than half on paper*	2	0
• *All on paper*	1	0
Did you use numbers instead of names to identify students and/or teachers on the forms and labels due to data protection/ confidentiality laws or rules in your country?	7	12

Note: *N* = 19.

Except for specific cases, such as Canada's decision to sample only five teachers per school, the only reported sampling-related concern was the reluctance of some sampled schools to administer the study to a sample of students *across* Grade 8 classes rather than to a sample based on intact classes. The NRCs from the Netherlands, Norway, and Switzerland stated that some schools in their countries were reluctant to comply with this requirement. The Netherlands and Switzerland NRCs said this reluctance was addressed by sampling intact classes in the few cases where sampling across the grade was a concern. In the case of Poland, variations to the sampling forms were an outcome of requirements regarding personal data protection.

In order to reduce task burden on the participating schools, some NRCs either hired external staff (Chile) or helped the schools prepare and complete the listing and tracking forms (Norway).

The second part of the SAQ collected information on contacting schools and recruiting school coordinators. Table 9.2 provides a summary of the responses to this section.

Table 9.2: Survey activities questionnaire responses: contacting schools and recruiting school coordinators

Question	Yes	No
Did you have any difficulties in convincing schools to participate?	12	7
How did you train the school coordinators?		
• *Formal training sessions*	8	–
• *Through telephone, email, or video-link*	11	–
• *Written instructions*	17	–
• *Other (please specify)*	2	–
Did the school coordinators report difficulties with understanding any of the following aspects of ICILS administration?		
• *Identifying eligible teacher and/or students*	4	15
• *The necessity for listing all target-grade teachers*	4	15
• *The necessity for listing all target-grade students*	6	13
• *Flagging students to be excluded prior to school sampling*	7	12
• *The rationale for sampling students from the target-grade across classrooms*	4	15
• *The processes for running the USB compatibility tests in schools*	4	15
• *The steps to set up computers in schools to support successful test administration*	5	14
• *The test administration procedures*	4	15

Note: *N* = 19.

In general, NRCs described contacting schools and recruiting school coordinators as a rigorous job. Between them, the NRCs used all means of contacting the sampled schools. The most frequent approach was by email (17 of the 19 responding NRCs) or phone (14 of the 19 NRCs). In some countries, ministries of education supported this process either by contacting principals (Canada) or sending them official letters (Lithuania, Turkey). In other countries, governments made it mandatory for selected schools to participate (Australia, the Czech Republic, and Korea).

Once a school had agreed to participate in ICILS, the school principal was usually the person who selected the school coordinator, with selection based on NRC-provided advice. The appointed school coordinators received training, either in person (eight NRCs) or through distance/online methods (11 NRCs). In most cases, coordinators also received the school coordinator manual or other written instructions produced by the NRCs.

This support equipped school coordinators well, although a minority of NRCs reported a few difficulties, the most frequently stated of which were school coordinators struggling to understand not only why the student sampling was across all classes of the target grade but also the rules for student exclusion prior to sampling.

NRCs also reported difficulties in achieving high participation rates of students, teachers, ICT coordinators, and school principals within the participating schools: nine countries (students), 10 countries (teachers), eight countries (ICT coordinators), and nine countries (principals).

Table 9.3 presents a summary of responses for questions related to adapting and translating the ICILS assessment materials.

Most countries translated the instruments directly from the English version and in accordance with their previous adaptations and translations from the field test. While NRCs did the national adaptations (or at least closely monitored them), external translators (editors, language correctors) produced the national versions of the instruments.

Table 9.3: Survey activities questionnaire responses: adapting and translating the ICILS assessment materials

Question	Adaptation	Translation
Who adapted and translated the international version of the student test modules?		
a) Own staff	19	16
b) Outside translator(s)	2	8
c) Outside reviewer(s)	2	2
Who adapted and translated the international version of the student, teacher, ICT coordinator and principal questionnaires?		
d) Own staff	19	16
e) Outside translator(s)	2	9
f) Outside reviewer(s)	1	2

Note: *N* = 19.

Overall, NRCs did not see the adaptation and translation work as difficult. However, there were instances where the ICT terminology (Chile and Korea) as well as some of the informal wording/vocabulary used in the student test module (Norway) proved difficult. Chile and Switzerland found adapting some questionnaires and the school coordinator manual challenging. Two countries (Germany and Norway) used only the English source version of the scoring guides.

The NRCs reported that assembling and preparing the ICILS materials for administration was not difficult, and they said they experienced no errors in the printing process. The only difficulties encountered (reported by three NRCs) related to the time taken to create copies of the USB sticks for use in the testing.

The third section of the SAQ asked NRCs to respond to questions relating to administering the student instruments and contextual questionnaires. A summary of responses to this section is presented in Table 9.4.

Table 9.4: Survey activities questionnaire responses: administering the student instruments and contextual questionnaires

Question	Number of cases
Who were the test administrators for the ICILS main survey?	
• *National center staff*	9
• *Regional or district government staff*	0
• *External contractor staff*	4
• *Teachers from other schools*	0
• *Teachers from the sampled schools but not of the sampled students*	11
• *Teachers of the sampled students*	4
• *School coordinators*	9
How did you train the test administrators?	
• *Formal training sessions*	14
• *Through telephone, email, or video-link*	9
• *Written instructions (in addition to the test administrator manual)*	9
• *Other*	6

Note: *N* = 19.

The majority of the test administrators were teachers from the sampled schools (15 NRCs). Most of these individuals also served as school coordinators. NRCs supported the ICILS test administration through their own involvement when necessary as well as by contracting external test administrators for this task.

Considerable attention was paid to training the test administrators. In most instances, training consisted of a formal (face to face) training session complemented by written instructions and the test administrator manual. Some countries developed additional means of training for their test administrators. Australia, for example, provided six online training modules, which were followed by a test and concluding teleconference.

During the test administration period, NRCs used the IEA SurveySystem monitor software application to check participation and the return status of online instruments. Five NRCs said they did this every day, 10 NRCs did so at least once a week, and only four NRCs said they checked less frequently or never. All NRCs who used this application found it useful.

Fifteen NRCs also organized monitoring of data-collection quality at the national level and visited a selection of the participating schools to observe test administration and interview school coordinators. In all of these cases but one, NRCs used the national quality control monitor manual template provided by the ICILS international study center (ISC). Some NRCs arranged for the manual to be translated or further elaborated to include procedures customized to the particular national level. On average, the participating countries appointed seven national quality control (NQC) monitors apiece, with the range extending from one (in Hong Kong SAR, Lithuania, and Turkey) to 16 in Poland and 22 in Chile. In Chile, the NQC monitors and the test administrators were trained together.

Across countries, the NQC monitors visited more than 10 percent of the schools that participated in the data collection. Seven NRCs reported undertaking some action as a consequence of the reported results (such as providing the test administrators with a newsletter for reporting outcomes of the quality control monitoring).

Four NRCs omitted the NQC monitoring procedure (one of these four countries cited a lack of resources as the reason). Two of the NRCs relied on IQCM conducted by IEA; one relied on experienced NRC staff to administer the assessment.

Quality control observations of the ICILS data collection

In order to ensure the high quality of the ICILS data collection activities, the international ICILS research consortium developed a quality control program designed to document how well countries complied with the international data-collection procedures. This program was an essential means of ensuring that all participating countries took a rigorous, standardized approach to data collection.

The IEA Secretariat hired and trained all IQC monitors in person. The training was supplemented by the international quality control manual, which contained all necessary information on the ICILS study framework and instruments, required international procedures, and IQC monitor roles and responsibilities. These roles and responsibilities included visiting NRCs, selecting schools for observation, commenting on the translation verification reports, establishing procedures for visiting the selected schools, and gathering up all required materials, including the rules for the financial compensation of the IQC monitors, and returning them to the IEA Secretariat.

The IEA Secretariat provided all NRCs with the national quality control monitor manual and asked them to manage the national quality control monitoring process. (The previous section of this chapter summarized some of the feedback from this activity.)

One month before the start of the main data collection in a country, each IQC monitor contacted the NRC to schedule a meeting with him or her. During this meeting, these individuals collected the necessary documentation and materials. They also selected (mostly through convenience sampling) the schools in which the monitors would conduct their observations, and obtained contact details for these schools.

The IQC monitors received materials and supplies from two sources—the IEA Secretariat and the NRC in their country.

The IEA Secretariat provided the following:

- *The international quality control monitor manual*, which explained the procedures IQC monitors were to follow;
- The international source version of the *school coordinator manual* and the *test administrator manual* in English;
- A *testing session observation tracking form*;
- An example of the *student listing form*, the s*tudent tracking form(s)*, the *teacher listing form*, and the *teacher tracking form*;
- Files containing the translation verification feedback and a *translation verification report*; and
- The *testing session observation record*.

The NRCs provided the following resources:

- A paper copy of the *school coordinator manual* for their country;
- A paper copy of the *test administrator manual(s)* for their country;
- A *student listing form* for each school selected for observation;
- A *student tracking form* for each school selected for observation;
- A *teacher listing form* for each school selected for observation;
- A *teacher tracking form* for each school selected for observation;
- A paper copy (or PDF file) of the *teacher questionnaire* (if the test was prepared in more than one language, one per language);
- A paper copy (or PDF file) of the *principal questionnaire* (if the test was prepared in more than one language, one per language);
- A paper copy (or PDF file) of the *ICT coordinator questionnaire* (if the test was prepared in more than one language, one per language); and
- A USB stick containing the student test and student questionnaire (if the test was prepared in more than one language, all languages available on one USB stick).

Before visiting schools, the IQC monitors reviewed the survey instruments and the translation verification outputs and used the translation verification reports to record their observations. This task had two aims: the first was to enable the IQC monitors to familiarize themselves with the study instruments and then to record any issues related to that content; the second was to let the IQC monitors check if the NRC-provided documentation on any adaptations implemented in the instruments correctly mirrored the situation in the final version of instruments used during the data collection.

Translation verification reports

The IEA Secretariat required all NRCs to submit all translated test materials for independent translation verification and to do this work before preparation of the test USB sticks and printing of the questionnaires. On receiving the materials back from the verification process, NRCs considered the comments or suggestions the verifiers had made in the documents and then decided whether or not to adopt these recommendations.

The IEA Secretariat asked the IQC monitors to determine whether each NRC had adopted the verifier's suggestions. When conducting their review, the monitors had at hand the final set of questionnaires, a USB stick containing the student test modules, the student questionnaire, and the same documents at the stage of containing the verifier feedback. A document (a spreadsheet) pertaining to the student test modules and questionnaire contained all of the verifier's comments, the NRC's original translations, the verifier's suggested translations, and the final translations. Verifier comments were inserted directly into the MS Word versions of the teacher, principal, and ICT coordinator questionnaires via "Track Changes."

In summary, the materials the IQC monitors received in order to review each NRC's adoption of verifier comments were copies of the teacher questionnaire, school questionnaire, ICT coordinator questionnaire, and a MS Excel file containing the verifier's comments for the student test materials and the student questionnaire.

Verifier comments on each of the above-listed documents were listed as Severity Code 1 (major changes, mistakes, and translation errors), Severity Code 1? (dubious cases), Severity Code 2 (grammatical issues), and Severity Code 3 (suggestion for alternative). For more information on severity codes, please refer to Chapter 6 of this report.

The IQC monitors compared the text in the final test instruments against the verifier's comments. If the NRC had adopted the verifier's suggestions, the monitors confirmed this by placing a check mark next to the verifier's comments or in the column provided in the Excel file. If the NRC had not adopted the verifier's suggestions, monitors wrote "No" next to the comments.

The IQCM findings revealed that approximately 80 percent of verifier comments and suggestions were implemented as suggested. Eleven percent of the verifier interventions led to further edits and review by the respective NRC, and nine percent were rejected and reverted back to the national version the NRC had submitted for translation verification. Of the nine percent of verifier comments that were rejected, approximately 80 percent of these came from the NRC for Hong Kong SAR and concerned the characters to be used in Hong Kong's Simplified Chinese version.

The overall results from the IQCM review of the translation reports showed that the translation verification process helped and improved the quality of the instruments.

Testing session arrangements and settings

The major role of the IQC monitors was to document data-collection activities in countries. To accomplish this, the monitors visited 15 schools in their respective countries in order to observe the data collection and to record the compliance of the test administration with the prescribed procedures. During their visits, the monitors observed the testing session, including the work of the test administrator during it, and

used the testing session observation record to record their observations. The monitors then interviewed the school coordinator and delivered the obtained data online. Finally, they returned hard copies of the testing session observation tracking form, the student listing form, the student tracking form, the teacher listing form, and the teacher tracking form to the IEA Secretariat.

The data collected by four sections of the testing session observation record concerned the arrangement of the actual testing session (including the security of the study materials), the procedures used during the actual testing session, the IQC monitors' general impressions of how the testing session was arranged and conducted, and accounts of the monitors' interviews with the school coordinators. The remaining sections of this chapter are based on this collected information. It is important to note that two IQC monitors were not able to collect the data in full. The monitor from Denmark observed only 13 testing sessions instead of 15, and the monitor from the Russian Federation did not attend all of each testing session. These missing data represented up to three percent of the overall count.

Overall, 11 percent of the testing sessions were held in rooms with 10 or fewer computers available for the assessment. This situation was evident in more than a quarter of the observed testing sessions in Lithuania and the Slovak Republic, about one third in Turkey, more than a third in the Russian Federation, and almost half of all observed cases in the city of Buenos Aires. Thirty-nine percent of the sessions had 11 to 20 computers, and more than 20 computers were available in 49 percent of the testing sessions. When 20 or more computers were available, all sampled students from the school could be assessed during one single testing session, a situation that was observed in Croatia, Hong Kong SAR, the Netherlands, and Germany (where the NRC sent 20 laptops to each school for the selected students).

If there were fewer than 20 computers in the testing room, countries either tested two student groups in parallel or tested the second student group immediately after the first one. The parallel option was the one most often used in Lithuania (i.e., in a third of the observed testing sessions). The testing one group after the other option occurred in almost two thirds of the test administrations in Poland and Slovenia, almost half of the observed cases in Turkey, more than a third of the observed sessions in Thailand, and in a third of the sessions in the Slovak Republic.

The ISC advised countries on what size the computer screens used in the assessment should be. Monitoring showed that 85 percent of students worked on computers with the required screen size of 38.1 centimeters (15 inches) or larger. In nine countries, all students worked under this condition. These countries were Croatia, the Czech Republic, Korea, Lithuania, Poland, the Slovak Republic, Slovenia, Thailand, and the city of Buenos Aires. Fourteen percent of students worked on computers with a screen size of less than 38.1 centimeters (15 inches).[1] In Germany, however, all students completed the assessment on computers with a screen size less than 38.1 centimeters, as did one third of the students in both Norway and Switzerland.

Of the computers used, 73 percent were desktop computers (all observed cases in Canada, the Czech Republic, Hong Kong SAR, Thailand, Turkey, and the city of Buenos Aires), and 25 percent were portable computers connected to a power plug (all observed

1 The missing one percent of data for this question was caused by 13 percent of missing responses from the Russian IQCM report.

cases in Croatia, Germany, and Poland). Only two percent were portable computers not connected to a power plug. This last group of computers was rarely used in Norway, the Russian Federation, Denmark, the Netherlands, and Australia.

In 89 percent of observed cases, a USB-based method was used to deliver the student instruments. These cases included USBs connected to laptops given to the schools specifically for use in the testing. A laptop server connected to the school network was evident in only five percent of administrations. This form of delivery was the preference in Lithuania (all but two observed cases) but rarely occurred in Australia and Switzerland (only one case each). Five percent of test administrations were delivered directly off a standardized model of laptop; all these cases were located in Poland. The missing observations from the Russian Federation accounted for the remaining one percent (where all 11 observed administrations used a USB-based delivery method.

There was adequate seating space for the students to work without distractions in 94 percent of cases overall. The remaining five percent was associated with a narrow space resulting in students being too close to one another and so sitting uncomfortably and/or being able to see the screens of their schoolmates. Inadequate seating space occurred in Poland (four observed administrations), Chile (three observed cases), Switzerland, Thailand, and the Russian Federation (two observed cases each), and the Slovak Republic and Turkey (one testing session apiece).

In most cases (97%), the test administrator had adequate room to move around during the test to ensure that students were following directions correctly. The two percent of cases where the test administrators experienced space limitations occurred in Hong Kong SAR, Chile, the Slovak Republic, Switzerland, and Thailand. However, these occurrences were rare in all five of these countries. All test administrators had a watch or used some other means to keep track of time during the testing session, except in one occurrence in the Slovak Republic and one in Slovenia.

Test administrators were expected to set up all workstations (either desktop computers or laptops) so that the test screen was visible before the students' arrival. Overall, 89 percent of the observed test administrators complied with this request, although 100 percent did so in Chile, Croatia, Germany, Korea, Lithuania, Norway, Poland, the Slovak Republic, and Slovenia. In eight percent of the observed administrations, test administrators either waited for the scheduled room to come available for the testing or faced some type of technical obstacle. In these cases, students arrived before the workstations had been installed.

In some instances, test administrators needed to install the assessment on the students' own notebooks, which meant they were unable to comply with the required procedure. The technical issues were evident only occasionally in Thailand (six testing sessions), Hong Kong SAR (four sessions), Australia (three sessions), and Switzerland, Turkey, and the Russian Federation (two observed cases each). Denmark, Canada, the Czech Republic, and the Netherlands each recorded only one occurrence.

The Russian Federation and the city of Buenos Aires accounted for the three percent of missing observational information across the participating countries. (In the Russian Federation, monitors failed to observe the entire test administration in 53% of cases.)

Most countries had each student's name displayed on the ICILS-assessment welcome screen. Overall, names were evident in 76 percent of cases, and in 100 percent of cases

in Australia, Canada, Croatia, Korea, Chile, Slovenia, and Switzerland. In 21 percent of cases, only student IDs were displayed on the login screen; Germany and Turkey used only student IDs in all cases. Student ID was the preferred display in Poland (almost two thirds of the observed administrations), and likewise in all eight documented cases in the Russian Federation.

In general, students were seated in the order in which their names appeared on the student tracking form (85% of cases), thus aligning with the ICILS recommendation. Canada, Croatia, Germany, Korea, Poland, Chile, Slovenia, and the city of Buenos Aires fully complied with this recommendation. Thirteen percent of test administrators or school coordinators arranged for a different order based on customs within the schools (e.g., having students seated at their usual seats in a computer lab or relocating talkative students). The number of such cases was relatively high in the Czech Republic (a third of the observed test administrations), Hong Kong SAR (almost half of the observed cases), and Turkey (more than half of the observed cases).

Test administrator activities and summary observations of test administration

One of the key pieces of information the IQC monitors gathered related to how the ICILS instruments were administered and the extent to which the intended timing of the test administration was followed. Tables 9.5 and 9.6 present an overview of this information. Because the IQCM reports provided some further description and explanation of the recorded deviations, additional details appear after the tables.

Table 9.5: Review of adherence to the timing described in the test administrator manual

	Yes %	No %	Missing %
Preparation of students, reading of instructions, and administering the tutorial	84	13	3
Administering the *student assessment, first module*	85	12	3
Short break	74	23	3
Administering the *student assessment, second module*	85	13	3
Short break	67	31	3
Administering the *student questionnaire*	81	16	3

Note: *N* = 298.

Table 9.6: Review of adherence to the procedures described in the test administrator manual

	Yes %	No %	Missing %
Allocating students to computers	90	7	3
Introducing the testing	91	6	3
Conducting the tutorial	89	8	3
Introducing the first assessment module	93	4	3
Conducting the first assessment module	93	4	3
Break arrangement	83	14	3
Conducting the second assessment module	93	5	2
Break arrangement	79	19	2
Conducting the questionnaire session	93	5	2
Releasing the students	95	3	2

Note: *N* = 298.

Differences in the reading of instructions and administration of the tutorial resulted in the reported changes in the time plan. The latter was a product of students needing more time than was allocated for the tutorial.

The monitors in seven countries reported no deviations in the time set down for completion of the test module(s). The countries were Canada, Croatia, Denmark, Germany, the Netherlands, Slovenia, and Turkey. Of the remaining countries, Switzerland experienced the most deviations (almost half of the observed cases), while Hong Kong SAR and Poland experienced deviations in more than one third of the observed test administrations. These deviations were either because of technical obstacles that meant the students affected needed extra time to complete their assessment or because all involved students completed the assigned modules sooner than expected (planned). These two reasons were the reasons most often given for a time deviation.

Major changes to the scheduled times involved breaks in the testing sessions. In some cases these were extended, shortened, or skipped altogether in order to maintain the school timetable or because the students wanted these changes. Test administrators were asked to allow additional time so as to support students in completing the questionnaire. Sixteen percent of time changes during the observed sessions involved giving students the necessary extra time (ranging from 5 to 18 minutes).

ICILS required the test administrators to follow the instructions given in the test administrator manual. The IQC monitors recorded whether or not they did so. According to the monitors' reports, administrators followed the prescribed procedures in most cases and in some countries in all cases. The most commonly occurring deviations related to the tutorial. Here, some test administrators modified the script by repeating, omitting, or rewording some parts. Other test administrators left students to explore the tutorial by themselves.

Monitors also recorded some issues with student behavior. The monitors for Thailand recorded the most deviations (a third of the observed cases), while those in Poland and Switzerland observed their occurrence in more than a quarter of their cases.

The first assessment module was not administered in the prescribed way in six of the 15 observed administrations in Hong Kong SAR, where the monitors' records showed students completing their work about 10 minutes sooner than expected. This outcome, recorded in all observed cases, might possibly be because Hong Kong did not follow the break arrangements but instead required students to work through the assessment without a rest. The records for two other countries showed a somewhat higher than expected deviation in break arrangements. These countries were Switzerland (a third of the observed administrations with respect to the first break and more than two thirds with respect to the second) and Thailand (a third of the observed administrations in the case of the first break).

Observations of test administration

According to the IQC monitors' records, in 94 percent of the observed cases, test administrators familiarized themselves with the test administration procedure and script before testing commenced. In four percent of cases, the monitors said the test administrators probably or definitely did not engage in this familiarization process. In Thailand, the administrators in almost half of all observations neglected this step.

The monitors recorded no technical problems with logging in or using the USB sticks/laptop server in 74 percent of their observed administrations. Of the 23 percent of observations where technical problems occurred, the incidence of these problems varied considerably across the relevant countries. While only one case was reported in each of Australia, Germany, Korea, Lithuania, and the city of Buenos Aires, the monitor in Hong Kong SAR reported technical problems in all but two of that country's administrations (these mostly related to inputting Chinese characters via the keyboard). More than half of the observed administrations in both Switzerland and Turkey featured technical problems, although no single issue provided a reason for them. The technical issues did not, however, influence the overall quality of the achievement data because they mainly arose before the test began and were quickly resolved. Chapter 3 contains details of the data collection across countries, while Chapter 11 contains information on the treatment of test-item data.

In 76 percent of cases, school information technology (IT) staff members were available for support during the whole testing session. This support was not available in 22 percent of administrations. In some countries, test administrators had to deal with problems by themselves much more often than their colleagues had to in the other countries. This was the situation in 12 out of the 15 observed cases in the Netherlands, 10 of the cases in Thailand, nine in the Slovak Republic, and eight in Turkey (thus more than half of the observed cases in each of these countries).

According to the IQC monitors' reports, students had problems with the testing in only 14 percent of the observed cases. The countries most affected were Turkey (more than two thirds of the observed cases), Chile and Lithuania (one third of the observed administrations each), and Switzerland and Hong Kong SAR (three documented cases). Turkey reported the tests as too difficult and confusing for students. Lithuania said students with special educational needs found it difficult to read and respond to the test as requested. In Chile, the test was reported to be too tiring, and in Hong Kong SAR and Switzerland students were adversely affected by the above reported technical issues.

Test administrators were able to provide technical support and respond to student needs appropriately in almost all cases—96 percent overall. Only two percent of the reported cases featured test administrators who did not address students' questions appropriately. These instances were mostly reported in Thailand (three occurrences). Although the test administrators in Thailand could solve the technical issues, they did not follow the script in the manual and (to use one monitor's words) "Students seemed to be confused about the test procedures."

The student questionnaire was mostly administered during the testing session (95% of all observations). Three percent of other cases occurred in Lithuania (almost half of the observed testing sessions were arranged this way in this country) and one documented case also occurred in Denmark. In 23 percent of cases overall, students asked for additional time to complete the questionnaire. This request mainly occurred in the Czech Republic, Slovenia, and Thailand (at least two thirds of the observed cases in each of these countries), and added six minutes on average to the testing time (nine minutes in the Czech Republic and seven minutes in Thailand).

In 58 percent of cases, students had queries about the questions on parental occupation in the student questionnaire. These queries arose during all administrations in Croatia. The countries where such queries least occurred were Germany (one case) and the Netherlands (two instances).

In 93 percent of the observations across countries, IQC monitors described students as being "somewhat orderly and cooperative." The countries where students were most noted as "hardly cooperative at all" were Switzerland and Thailand (two records each). In both cases, the test administrators made at least "some effort" to control the students and the situation. In 83 percent of the observed testing sessions, the monitors saw no evidence of students attempting to cheat on the test or not paying attention to it. The highest rate of observed disruptive behavior occurred in Switzerland (almost two thirds of observations) and in Thailand and Hong Kong SAR (about half of the observed administrations each). Most of the recorded cases involved students talking to one another, but the monitors saw no evidence during these incidents of an intention to cheat.

The IQC monitors rated the overall quality of the testing session as "excellent" in 49 percent of the observations (all but two cases in Australia and in Germany), "very good" in 28 percent of the observations (almost two thirds of observed cases in Turkey), "good" in 16 percent of the observations (six cases in Poland), and "fair" in five percent of the observations (three cases each in Thailand and in Russia). Of all the testing sessions, only one percent received the evaluation of "poor." This was in Thailand, where two such occurrences were observed.

Interviews with school coordinators

Forty-five percent of the interviewed school coordinators were teachers from the participating schools. In Thailand and Turkey, all school coordinators were teachers. In Hong Kong SAR, the Slovak Republic, and Slovenia, the percentage of teachers who assumed the school coordinator role was very high: only two or three of the 15 interviewed school coordinators were not teachers in these countries. Thirty-seven percent of the school coordinators belonged to school management teams. In Australia and Norway, all but two coordinators were not teachers, in the Russian Federation the number was three, and in the Netherlands four. In Croatia, school librarians and psychologists served as coordinators (in all but three recorded exceptions). External staff performed this role in Chile and Switzerland.

Most school coordinators (89%) were satisfied with the school coordinator manual, stating in response to the relevant question that it "worked well." However, 10 percent noted that the manual needed some improvement. The number of occurrences of coordinators stating the need for improvements ranged from none in the Czech Republic, Norway, the Slovak Republic, and the Russian Federation up to four in Canada as well as in Lithuania and nine in Switzerland. School coordinators in Canada and Lithuania generally considered the manual too wordy, while several of the Swiss school coordinators were not happy with the quality of the German translation or required more advice on how to deal with particular technical problems.

According to the school coordinators, the testing went very well, with no problems recorded in 73 percent of schools across the participating countries. Coordinators in Turkey recorded problems in four cases, while Switzerland recorded problems in just under half of all cases. Testing progressed satisfactorily with only a few problems in 27 percent of schools. Coordinators rated the attitude of other school staff members toward the ICILS testing as positive in 59 percent of cases, neutral in 37 percent of cases, and negative in two percent of cases. The highest percentages reporting an indifferent (neither positive nor negative) attitude were in Australia (all but three cases), Canada (all but four records), and the Czech Republic (two thirds of all records).

The school coordinators confirmed the correct shipment of all the necessary items: school coordinator manuals, test administrator manuals, the USB sticks (when applicable), teacher questionnaires (or cover letters), principal questionnaires (and cover letters), ICT coordinator questionnaires (or cover letters), and the teacher tracking form. The coordinators in the city of Buenos Aires (two recorded cases) reported minor deviations in shipment of these materials.

The school administrators verified adequate supplies of the USB sticks and that the testing materials were safely stored before the test was administered. Lithuania did not use USB sticks because it administered the test through school networks. In Croatia, Hong Kong SAR, and Poland, the test administrators, who were external to the school, brought the USB sticks with them to the testing sessions.

ICILS expected test administrators to run the USB comparability test on computers before the assessment began. Overall, this did not happen in 18 percent of cases. However, in all these occurrences, the decision was justified for reasons such as the computers being brought to school, tested during the field-test stage, or checked by another person (e.g., ICT coordinator). The arrangements schools made to accommodate the testing session were mostly seen as satisfactory (97%). Denmark recorded two deviations in this regard. Here, the difficulties noted included having to reschedule other classes to accommodate the test and having to arrange the testing during school hours.

Eighty-three percent of the test administrators thought that makeup sessions would not be required at their school. However, in Switzerland, administrators voiced this requirement in all but three cases. Four such cases were reported in Chile and Norway. All of these administrators were prepared to conduct a makeup session if needed.

Sixty percent of IQCM records reported that the sampled students received special instructions, a motivational talk, or incentives to prepare them for the assessment. This happened in all observed cases in Australia, Korea, the Slovak Republic, Turkey, and the Russian Federation. In most instances, these talks were explanatory and focused on the content of the study and testing arrangements (e.g., "Students were briefed on the assessment purpose, but not motivated"). In Korea, the national research center provided all sampled students with a small gift. In many occurrences, parents were given the introductory information on the assessment. In 38 percent of the administrations, students received no spoken instructions or motivational statements before they began the assessment. This situation was prevalent in Germany (all observed cases), Denmark (all but one), Thailand (all but two), the Czech Republic (all but three), and Canada and Poland (all but four each). In one percent of the observations, the pre-assessment information was missing.

Other information

All of the IQC monitors except one stated that students from all classes of the ICILS target grade in the visited school were recorded on the student listing form, which meant the students who were expected to attend the assessment did so. The exception was the city of Buenos Aires, where the missing records were caused by the omission of one or more teaching shifts (morning, afternoon, and/or evening one). No additional students attended the observed testing sessions.

In the case of teachers, the IQC monitors noted five percent of cases where the number of teachers appearing on the teacher listing form was not the same as the number of

teachers present in the classroom schedule for the ICILS target grade. However, there were only two observations (out of the total 298 realized school coordinator interviews) where IQC monitors found that the school coordinator used either a convenience sample or listed the most active teachers teaching at the target grade. In all other cases, the additional teachers were new or replacement teachers.

Overall, the teacher questionnaires or the cover letters for teachers were distributed exactly according to the teacher tracking form. However, in six percent of cases this distribution was not done before the student testing session began. In Switzerland (almost half of the recorded cases) and the Netherlands (three cases), the test administrators rather than the school coordinators distributed the teacher questionnaires/cover letters. Twenty-seven percent of the interviewed school coordinators reported that the teacher questionnaires were not completed prior to the test administration (more than three-quarters of all observed cases in the Netherlands, almost two thirds in Lithuania, and more than half in Switzerland).

In the case of Lithuania, teachers filled in online questionnaires, and their response window was customized for their convenience. The situation with respect to the principal questionnaires was very similar. Overall, 24 percent of cases featured noncompleted principal questionnaires. This situation was evident in more than two thirds of the cases in the Netherlands and in almost two thirds of the cases in Switzerland.

Considerable effort was made to keep the administered questionnaires short and thereby lower the burden on respondents. The ICILS research team expected that the teacher questionnaire would take about 30 minutes to complete. That estimate was correct, according to the school coordinators' experiences, in 54 percent of cases (in all cases in Germany and Turkey, all but one in Poland and in the Russian Federation, and all but two in Korea).

Twenty-eight percent of the records indicated that teachers spent less time (5 to 15 minutes less) than anticipated on the teacher questionnaire. This was mostly the case in Croatia (more than two thirds of the recorded answers), Slovenia and Thailand (almost two thirds), and the Czech Republic and Switzerland (more than half of the records). In contrast, teachers spent more time than was expected on the questionnaires in only two percent of the cases, with the time differential again ranging from 5 to 15 minutes. This situation was reported in four countries: Denmark, Chile, and Hong Kong SAR (two records each), and the Slovak Republic (one record). Sixteen percent of the interviewed school coordinators did not answer the question relating to time.

The ICILS research team expected that the principal questionnaire would take principals about 20 minutes to complete. Here, data on how much time principals did spend was missing in a good number of cases; in 24 percent of the school coordinator interviews, the coordinators said they did not know the answer to this question. In 63 percent of cases the estimate of 20 minutes proved to be correct: all recorded cases in Korea, the Russian Federation, and Turkey, all but one in Germany, and all but two in the Czech Republic and Poland. In 11 percent of cases, school coordinators claimed that they took no more than 20 minutes to fill in the principal questionnaire. This was the situation in almost half of the cases in Switzerland and more than one third in Slovenia. Only two percent of the school coordinator interviews revealed instances of principals taking more time than anticipated to complete their questionnaire; the two cases recorded in Thailand was the maximum number of such cases.

Summary

Quality assurance was a very important part of ICILS. Quality assurance procedures covered instrument development, sampling, translation, verification of the national versions of all instruments, field operations, and data collection, scaling, analysis, and reporting. Perspectives on the integrity of the implementation of ICILS were derived from the survey activities questionnaire (SAQ) and the international quality control monitoring (IQCM).

The SAQ was completed by national research coordinators (NRCs), and the IQCM was conducted by independent monitors trained by the IEA Secretariat. IQC monitors visited 15 of the participating schools in each country on the day of the test to observe the administration sessions and to interview the school coordinators. Overall, monitors concluded that ICILS was implemented in ways that complied with the study's intended design. Observed deviations from expected procedures and practices were small in number, and any implementation issues were rectified promptly without any impact on data quality.

CHAPTER 10:

Data Management and Creation of the ICILS Database

Michael Jung and Ralph Carstens

Introduction

This chapter describes the procedures for checking ICILS data and for database creation that were implemented by IEA's Data Processing and Research Center (IEA DPC), the ICILS international study center (ISC) at the Australian Council for Educational Research (ACER), and the national centers of the participating countries. The main goals of these procedures 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 in fact 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 data.

Data sources

Computer-based student assessment

As a computer-based assessment, ICILS 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 four total test modules as well as the student questionnaire. The national centers provided schools with one USB stick per student (plus spare sticks in case of technical failure) or laptop computers if the computers at a school were deemed unsuitable for administering the student assessment. 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 most cases, the data from the student assessment were stored on individual USB sticks because this was the default administration method in ICILS. If the server method was utilized, then the student data were instead 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. Accordingly, during this step a set of calculations needed to take place. Examples included automatic scoring of some of the complex or constructed-response items or aggregating time spent on an item across multiple visits to it.

Online data collection of school and teacher questionnaires

ICILS offered online collection of school and teacher questionnaire data as an international option conducted according to a mixed-mode design. Participating countries could adopt the online option as a default data-collection mode for some or 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.

As respondents completed their online questionnaires, their data were automatically stored in one 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 the IEA DPC) with the data from the paper-based questionnaires and also 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 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 any paper-based principal questionnaires, ICT coordinator questionnaires, and teacher questionnaires into computer data files using the IEA Data Management Expert (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.

Before data entry commenced, national centers were required to adapt the codebook structure to reflect any approved adaptations made to the national questionnaire versions (e.g., a nationally added response category). The IEA DPC verified these adapted codebooks, which then served as templates for creating the corresponding dataset.

In general, national centers were instructed to discard any questionnaires that were unused or that were returned completely empty, but 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 DME required national staff to enter data "as is" without any interpretation, correction, truncation, imputation, or cleaning. Resolution of any inconsistencies remaining after this data-entry stage was delayed until 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 "check-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 was outside the originally expected range (e.g., if an ICT coordinator reported more than 1,000 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 was logically inconsistent.
- If responses were not given at all, were not given in the expected format, were 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 DME were automatically validated. First, the entered respondent ID was validated with a three-digit code—the checksum, generated by the Within-School Sampling Software (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 DME also enabled identification of a range of problems such as inconsistencies in identification 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

It was expected that most countries would collect most of the data for the principal questionnaire, ICT coordinator questionnaire, and teacher questionnaire online. Paper data capture was therefore not expected to be as extensive as would obviously be the case in paper-only surveys.

To check the reliability of the paper data entry within participating countries, national centers were required to have two different staff members enter all (100%) of the completed principal, ICT coordinator, and teacher questionnaires. The IEA DPC recommended that countries 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 national center 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.

The national centers of only two countries, both of which had high proportions of paper-based questionnaires, contacted the IEA DPC for alternative suggestions and support regarding double data entry of a percentage or number of paper-based instruments.

While it was desirable that each and every discrepancy be resolved before submission of the complete dataset, the acceptable level of disagreement between the originally entered and double-entered data was established at 1.0 percent or less; any value above this level required a complete re-entry of 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 originally entered and double-entered data was evaluated by the IEA DPC. Apart from the nine countries that administered all of their contextual questionnaires online, data for another nine countries showed no differences between the main files and the files created for the purpose of double-data entry. In the remaining two countries that entered only a subset of records twice, the error rate was less than 0.1 percent in most and less than 1.0 percent in all datasets.

Data verification at the national centers

Before sending the data to the IEA DPC for further processing, national centers carried out mandatory validation and verification steps on all entered data and undertook corrections as necessary. The corresponding routines were included in the DME software, which automatically and systematically checked data files for duplicate identification 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 nonempty 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 WinW3S.

In addition to submitting the data files described above, national centers provided the IEA DPC 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. The IEA DPC 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, the data manager or other staff at 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 WinW3S.

This process was also supported by the option in WinW3S to generate an inconsistency report. This report listed all of the types of 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 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 the IEA DPC. If inconsistencies remained or the national center could not solve them, the DPC asked the center to provide documentation on these problems. The DPC used this documentation when processing the data at a later stage.

Confirming the integrity of the international and national databases

Overview

Ensuring the integrity of the international database required close cooperation between the international and national institutions involved in ICILS. Quality control comprised several steps. During the first step, staff at the IEA DPC checked the data files provided by each country. They then applied a set of (in total) 140 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 DPC staff sent queries to the national centers. These required the centers either to confirm the DPC's proposed data-editing actions or to provide additional information to resolve inconsistencies. After all modifications had been applied, staff at the IEA DPC 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, the IEA DPC provided national centers with univariate statistics at national level and international level. This material enabled national center staff to compare their national data with the international results included in the draft international report and with 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 indicate an error in translation, data capture, coding, etc. The international team reviewed all such instances and, when necessary, addressed it by, for example, 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, staff at the IEA DPC sent data to the ISC, which then produced and subsequently reviewed the basic item statistics. At the same time, the IEA DPC produced data files containing information on the participation status of schools, students, and teachers in each country's sample. Staff at the IEA DPC then used this information, together with data captured by the software designed to standardize operations and tasks and 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 computer and information literacy performance 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 the IEA DPC for inclusion in the international database and for distribution to the national centers.

Data cleaning quality control

Because ICILS was a large and highly complex study with very high standards for data quality, maintaining these standards required an extensive set of interrelated data-checking and data-cleaning procedures. To ensure that 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, the IEA DPC applied them to simulation datasets containing all possible problems and inconsistencies.
- *Registering all incoming data and documents in a specific database:* The IEA DPC recorded the date of arrival as well as specific issues requiring attention.
- *Carrying out data cleaning according to strict rules:* Deviations from the cleaning sequence were not possible, and the scope for involuntary changes to the cleaning procedures was minimal.
- *Documenting all systematic data recodings that applied to all countries:* The IEA DPC recorded these in the *ICILS General Cleaning Documentation* for the main survey (IEA, 2014).
- *Logging every "manual" correction to a country's data files in a recoding script:* Logging these changes, which occurred only occasionally, allowed IEA DPC staff 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 the IEA DPC 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:* The IEA DPC 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.

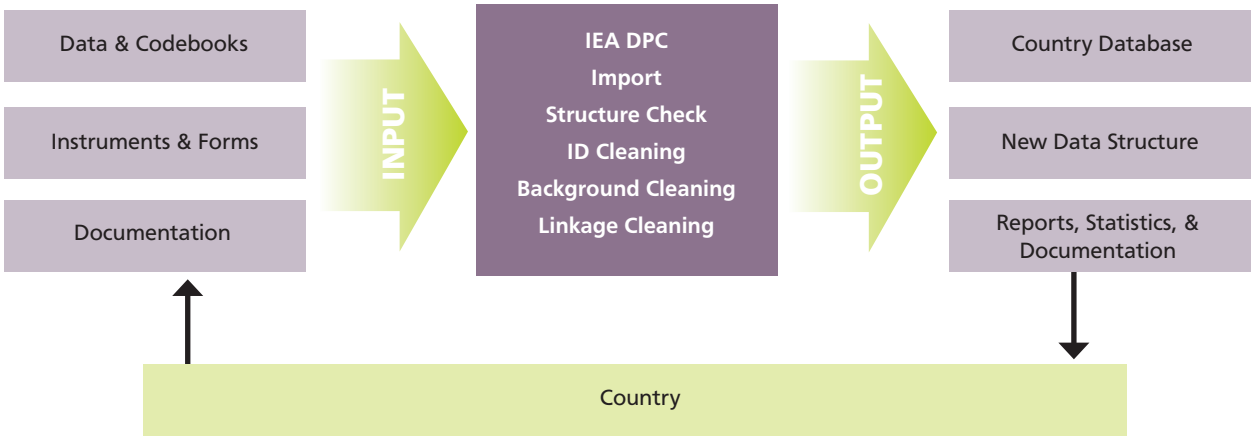
The IEA DPC compared national adaptations recorded in the documentation for the national datasets against the structure of the submitted national data files. IEA DPC staff then recorded any identified deviations from the international data structure in the national adaptation database and in the *ICILS User Guide for the International Database* (Jung & Carstens, 2015). Whenever possible, the IEA DPC recoded national deviations to ensure consistency with the international data structure. However, if international comparability could not be guaranteed, the IEA DPC 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 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). The IEA DPC carried out all of these activities in close communication with the national centers.

Figure 10.1: Overview of data processing at the IEA DPC



Checking documentation, import, and structure

For each country, data cleaning began with an exploratory review of its data-file structures and its data documentation (i.e., national adaptation forms, student tracking forms, teacher tracking forms, survey activities questionnaire).

The IEA DPC began data cleaning by combining the tracking information and sampling information captured in the WinW3S database with the student-level database containing the corresponding student survey instrument data. During this step, IEA DPC staff also merged 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 all further data-processing stages.

The first checks identified differences between the international and the 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 cultural or language-specific terms), whereas other countries inserted response categories within existing international variables or added national variables.

To keep track of adaptations, staff at the IEA DPC asked the national centers to complete national adaptation forms (NAFs) while they were adapting the international codebooks. Where necessary, the IEA DPC 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 User Guide for the International Database* (Jung & Carstens, 2015).

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

Once IEA DPC staff 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 the IEA DPC 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. The action the cleaning program or IEA DPC staff took with respect to each problem was also recorded.

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

Cleaning identification variables

Each record in a data file needs to have a unique identification number. The existence of records with duplicate ID numbers in a file implies an error of some kind. If two records in an ICILS database shared the same ID number and contained exactly the same data, the IEA DPC deleted one of the records and kept the other one in the database. If both records contained different data and IEA DPC staff found it impossible to identify which record contained the "true data," they removed both records from the database. The IEA DPC tried to keep such losses to a minimum; actual deletions were, in the end, very rare.

Although the ID cleaning covered all data from all instruments, it focused mainly on the student file. 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.

As mentioned earlier, the IEA DPC conducted all cleaning procedures in close cooperation with the national centers. After national center staff had cleaned the identification variables, they passed the clean databases with information about student participation and exclusion on to the IEA DPC sampling section, which used this information to calculate students' participation rates, exclusion rates, adjudication flags, and student sampling weights (see Chapter 7 for details).

Checking linkages

In ICILS, data about students, their schools, and teachers appeared in a number of different 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 values in the student main 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. However, none of the national datasets was completely free of inconsistent responses. The IEA DPC determined the treatment of inconsistent responses on a question-by-question basis, using all available documentation to make an informed decision. IEA DPC staff 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 possibly exceed the number provided by school principals in Question 3. Similarly, it was not logically possible for the sum of all fulltime teachers and parttime teachers, as asked in principal questionnaire Question 6, to equal zero.

In another example, Question 7 of the ICT coordinator questionnaire asked coordinators to provide the total number of computers in the school, the number of computers available to students, and the number of computers connected to the internet. Logically, the total number of computers in the school could not be smaller than the number available to students or connected to the internet.

The IEA DPC flagged inconsistencies of this kind and then asked the national centers to review these issues. IEA DPC staff recoded as "invalid" those cases that could not be corrected or where the response provided was not usable for analysis.

Filter questions, which appeared in some questionnaires, directed respondents to a particular subquestion or further section of the questionnaire. The IEA DPC applied the following cleaning rule to these filter questions and the dependent questions that followed: If the answer to the filter question was "no" or "not applicable," the IEA DPC recoded any responses to the dependent questions as "logically not applicable."

The IEA DPC also applied what are known as split variable checks to questions where the answer was coded into several variables. For example, Question 23 in the student questionnaire asked students: "At school, have you learned how to do the following tasks?" Student responses were captured in a set of eight variables, each one coded as "Yes" if the corresponding option was checked and "No" if the option was left unchecked. 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 data indicated age and gender for both teachers and students. The first set was the tracking information provided by the school coordinator or test administrator throughout the within-school sampling and test/questionnaire administration process. The second set comprised the 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, this information, if available, was copied over from the tracking data to the questionnaire .

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 existing 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 data capture: valid data values, and missing data values. Missing data can be assigned a value of omitted, invalid, or not administered during data capture. With the exception of the "not reached" missing codes assigned at ACER, the IEA DPC 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 noninterpretable or out of range.
- *Not administered:* This signified that the item or question was not administered to the respondent, which meant, of course, 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 (two out of four) administered to a student either deliberately (due to the rotation of modules) or, in a 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 to him or her.
- *Not reached* (this applied only to the individual items of the student test): This code indicated those items that 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 within the same *booklet*[1] following it (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.

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

Checking the interim data products

Building the international database was an iterative process. Once the IEA DPC completed each major data-processing step, it sent a new version of the data files to the national centers so that staff 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 database. All interim data were made available in full to the ISC at ACER, whereas, as just mentioned, each participating country received only its own data.

The IEA DPC sent the first version of data and accompanying documentation to countries in November 2013. At this time, data for 11 countries were sent out. The data for the remaining countries were sequentially provided as they became available, with this time period usually being six to eight weeks after countries had submitted their national data files to the DPC. This first version of each country dataset included the following data and documentation:

- School-, student- and teacher-level SPSS data files;
- Univariate descriptive statistics for all variables in the SPSS 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 the IEA DPC; and
- Cleaning documentation describing not only the initial cleaning procedures undertaken at the IEA DPC but also the data files and statistics provided.

The DPC provided NRCs and ACER with subsequent versions of the data and related documentation as soon as it had implemented feedback from countries. These additional versions of the data files were accompanied with the sampling weights and international achievement scores as soon as these became available. During this stage of the data-processing process, the DPC asked countries to review the documentation on adaptations to the national versions of their instruments and the related edits applied to the data files.

During the fifth NRC meeting in Copenhagen in June 2014 and for three weeks following it, NRCs had opportunity to raise any issues concerning their data that had thus far gone unnoticed. This step resulted in an updated data version that concluded the field work and included scale scores. The IEA DPC sent this version to the NRCs and to ACER in August 2014. These data meant that all countries could now replicate the results presented in the draft international reports.

In September 2014, NRCs received an update not only of their data but also of the data from all other countries. The update reflected minor issues that had been raised and resolved after the August 2014 data release. The ISC at ACER used this version of the data to produce the updated, final tables for the international report.

The ICILS international database

The ICILS 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 identification variables or removing some of the personal data variables that were needed only during field operations and data processing.

Once each NRC and the ISC had agreed on data-release policy and confidentiality agreements, the DPC made available a draft international database that included data from all participating countries. This step occurred in August 2014, prior to publication of the international report in November 2014. The release enabled countries to replicate the results presented in the draft chapters of the international report.

More information about the ICILS international database is provided in the *ICILS User Guide for the International Database* (Jung & Carstens, 2015).

Summary

To achieve high-quality data, ICILS implemented a series of data-management procedures that included checks to ensure the consistency of national database structures, provide proper documentation of all national adaptations, and safeguard the comparability of international variables across national datasets. Staff at the IEA DPC 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 creation of the final ICILS database. The final data products included item statistics, national data files, and the international database accompanied by a user guide and supplementary information.

References

International Association for the Evaluation of Educational Achievement (IEA). (2014). *ICILS general cleaning documentation*. Amsterdam, the Netherlands: Author.

Jung, M., & Carstens, R. (Eds.). (2015). *ICILS 2013 user guide for the international database*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

CHAPTER 11:

Scaling Procedures for ICILS Test Items

Eveline Gebhardt and Wolfram Schulz

Introduction

This chapter describes the procedures used to analyze and scale the ICILS test items that were administered to measure students' computer and information literacy (CIL). 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 crossnational measurement equivalence;
- International item adjudication;
- International item calibration and test reliability; and
- International ability estimates (plausible values and weighted likelihood estimates).

The development of the ICILS test items is described in Chapter 2 and was guided by the ICILS assessment framework (see Fraillon, Schulz, & Ainley, 2013).

The scaling model

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

Use of the one-parameter (Rasch) model (Rasch, 1960) for dichotomous items means that the probability of selecting Category 1 instead of 0 is modeled 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 is the estimated ability 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 (m_i) categories, we can generalize this model to the partial credit model (Masters & Wright, 1997), which takes the form of

$$P_{x_i}(\theta_n) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_i + \tau_{ki})}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} (\theta_n - \delta_i + \tau_{ki})} \quad x_i = 0, 1, \mathrm{K}, \ldots, m_i$$

Here, $P_{x_i}(\theta_n)$ denotes the probability of person n scoring x on item i, and θ_n denotes the person's ability. The item parameter δ_i gives the average location of the item on the latent continuum; τ_{ki} denotes the k^{th} step parameter for the multiple scores.

ACER Conquest, Version 3.0 software (Wu, Adams, Wilson, & Haldane, 2007) was used to scale the ICILS test data.

Test targeting

When measuring cognitive abilities, it is important to use test items that cover the different levels of achievement found in the target population. Figure 11.1 shows the distribution of cognitive abilities among ICILS students (information on the representative sample used for the final calibration appears later in this chapter). The figure also shows the location of items with the default response probability of 0.5 (indicating the point where, for example, a respondent is equally likely to provide a correct or an incorrect response to a multiple-choice item).

The range of item difficulties broadly matched the abilities found in the student population. However, the average item difficulty (0 logits) was somewhat higher than the average student ability (-0.05 logits). Therefore, the test items were somewhat better at describing the ability of students in the higher than in the lower score ranges of the CIL international achievement scale. However, the match between test item difficulty and student ability varied considerably across countries depending on the distribution of student achievement within each ICILS country.

Assessment of item fit

Before reviewing the international scale and more specific item statistics in detail, we used a set of different indicators to determine the goodness of fit for individual items and removed items with unsatisfactory scaling properties from subsequent analyses. Items can violate the assumptions of the IRT scaling model in a variety of ways. Therefore, reviewing item fit is based on a combination of assessments, such as mean square statistics, item-rest correlations, item characteristic curves, percentages of students in each response category, and the average ability of students in each response category.

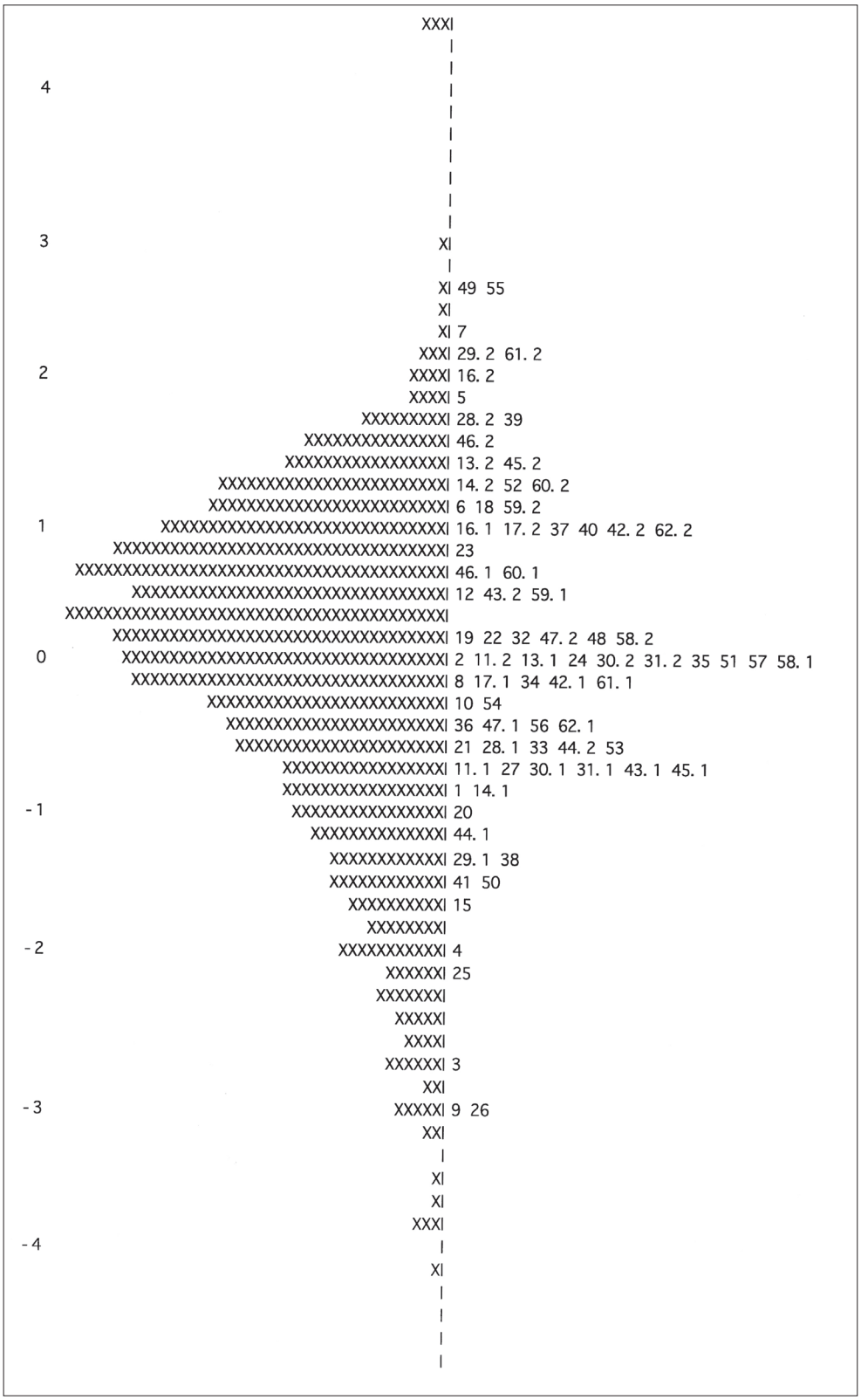
One way to determine goodness of fit is to calculate a (weighted) mean square statistic (Wright & Masters, 1982). Reviewing this residual-based item fit gives us an indication of the extent to which each item fits the item response model. However, there are no clear rules for acceptable item fit, and some statisticians recommend that analysts and researchers interpret residual-based statistics with caution (see, for example, Rost & von Davier, 1994). We kept these caveats in mind when assessing ICILS item fit by using a broad range of item statistics.

Item characteristic curves (ICC) provide a graphical representation of item fit across the range of student abilities for each item, including dichotomous and partial credit items. The horizontal axis represents the measured latent trait (here, students' CIL), while the vertical axis depicts the probability of obtaining the score.

As an example, Figure 11.2 shows the ICC for Item A03Z, a multiple-choice question with four response options. Observed curves (the broken lines in the figure)[1] were plotted for each response category together with the expected curve for the correct response (solid line), which follows the prediction of the Rasch model. The slope of the correct response increases and the slopes of the incorrect responses decrease. The item

1 This multiple-choice item had five categories: A, B, C, D, and 9 for missing. The broken line at the bottom of the graph pertains to all categories in the data.

Figure 11.1: Mapping of student abilities and item difficulties



Note: Each 'X' represents 12.1 cases.

difficulty parameter of -2.66 indicates that this item was a very easy one. The observed curve for the correct response is very close to the expected curve, which corresponds to a weighted mean square statistic close to one, thus indicating a good fit to the scaling model.

Figure 11.2: Item characteristic curve by category for Item A03Z

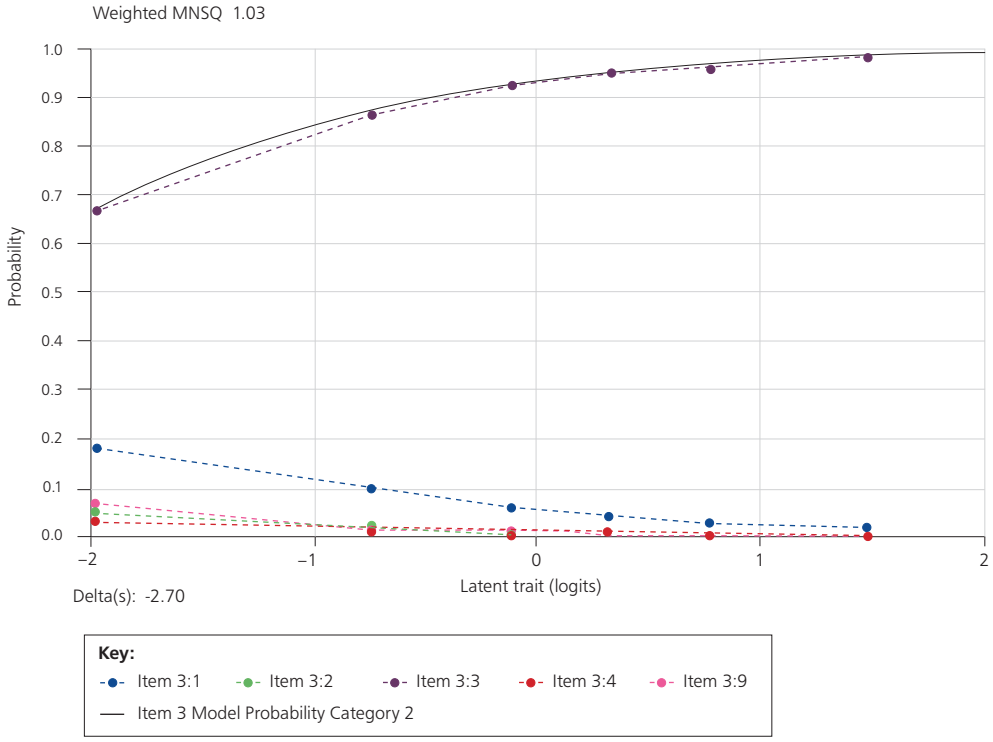


Figure 11.3 shows the ICC for Item B03Z, which is a dichotomously scored constructed-response item. The solid line represents the proportions of expected responses under the Rasch model, while the broken line indicates the observed proportions across ability groups. Here we can see that the observed curve fits the expected model very closely, an outcome that is also suggested by the weighted mean square statistic of 0.99. The item parameter of 0.24 indicates this item as one of medium difficulty.

The third ICC (in Figure 11.4) shows the observed and expected curve for each score for a partial credit item (A10D). The intersection between the curve for a score of 0 and the curve of the maximum score 2 corresponds to the location of the overall difficulty of this item (approximately 0.2 logits), thus indicating medium difficulty.

We further analyzed the functioning of the constructed-response scoring guides by reviewing the proportion of responses in each score and the correct ordering of mean abilities of students across scores. We also reviewed the (item-rest) correlations between the scored items and the corresponding total score based on all other items. We usually flagged item-rest correlations of 0.20 or lower for further review.

This analysis led to a set of scaled scored items that had satisfactory model fit in terms of the review of scaling properties at the international level. Five items were removed from scaling at this stage due to unsatisfactory scaling properties (A10K, B07A, H04A, S02Z, and S05Z). Two more items were deleted internationally at later stages because of differential item functioning, while two further items were merged into one item due to local dependency issues (see sections below for details).

Figure 11.3: Item characteristic curve by score for dichotomous Item B03Z

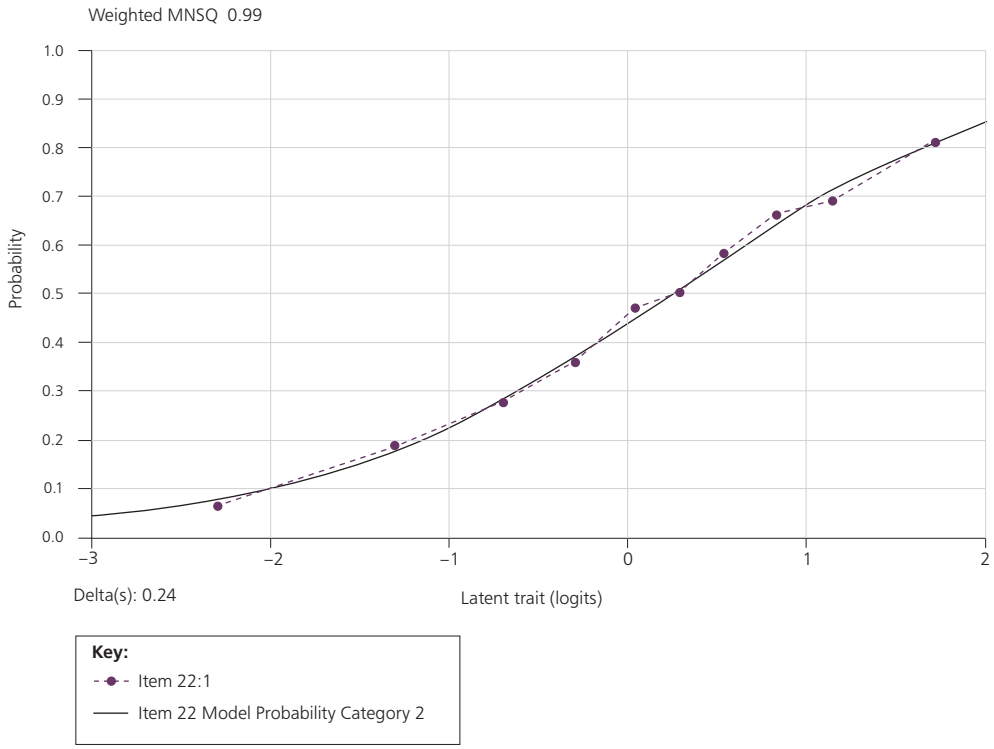


Figure 11.4: Item characteristic curve by score for partial credit Item A10D

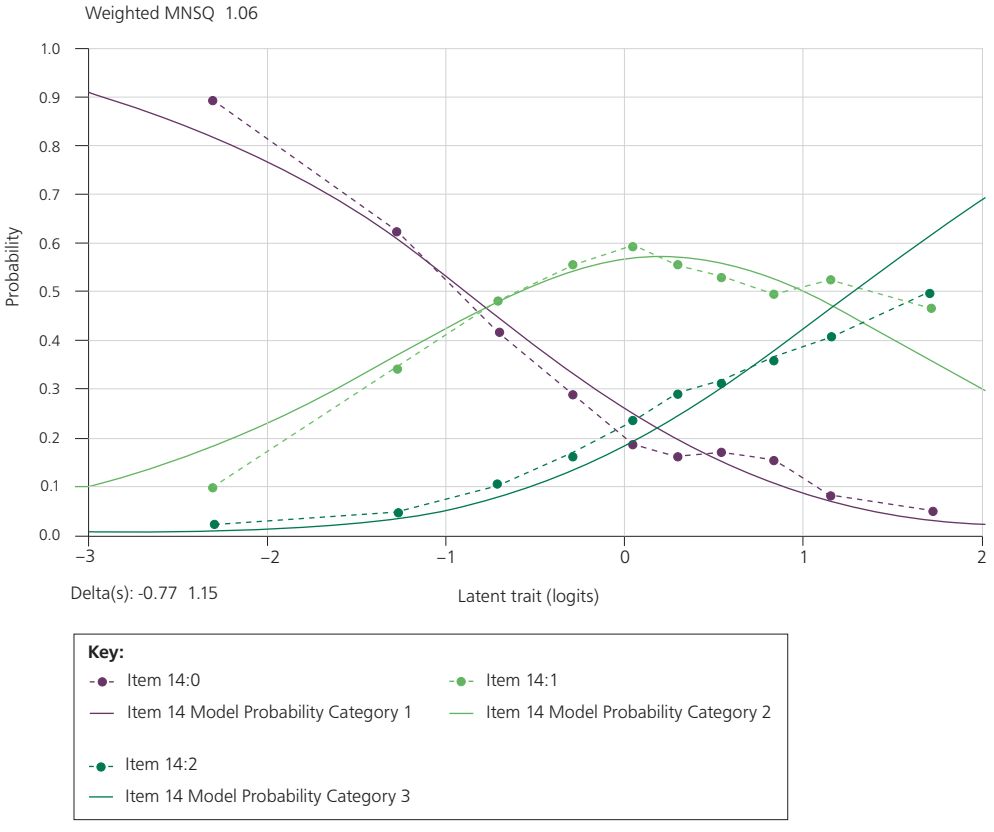


Table 11.1 shows the item-rest correlations[2] between correct responses to multiple-choice items or scored partial credit items and the weighted item fit statistics for the item set selected after the first review stage. The item-rest correlations for this item set ranged from 0.18 to 0.68. The two items with item-rest correlations below 0.20 were very difficult items (10% of correct responses on average). All other item characteristics, however, suggested satisfactory scaling properties for these items. The weighted mean square statistics for all items were below 1.20, thereby indicating satisfactory item discrimination between high-performing and low-performing students in ICILS.

Table 11.1: Item-rest correlation and weighted mean square statistics for ICILS test items

Item	Item-Rest Correlation	Weighted MNSQ	Item	Item-Rest Correlation	Weighted MNSQ
A01Z	0.43	1.00	H01Z	0.32	1.08
A02Z	0.40	1.03	H02Z	0.37	1.01
A03Z	0.31	1.03	H03Z	0.45	0.96
A05Z	0.34	1.05	H05Z	0.24	1.07
A06A	0.29	1.02	H06Z	0.35	1.03
A06B	0.35	1.04	H07A	0.57	0.82
A06C	0.24	1.02	H07B	0.60	0.93
A07Z	0.38	1.06	H07C	0.68	0.85
A08Z	0.35	0.91	H07D	0.51	1.12
A09Z	0.48	0.96	H07E	0.44	1.12
A10A	0.64	0.93	H07G	0.41	1.14
A10BF	0.38	1.04	H07H	0.58	1.01
A10C	0.59	0.94	H07I	0.45	0.97
A10D	0.49	1.06	H07J	0.20	1.03
A10E	0.40	0.92	S01Z	0.28	1.12
A10G	0.41	1.06	S03Z	0.26	1.16
A10H	0.59	0.96	S04A	0.32	1.03
A10I	0.39	0.99	S04B	0.35	1.08
A10J	0.48	0.94	S06Z	0.23	1.20
B01Z	0.36	1.07	S07Z	0.18	1.04
B02Z	0.42	1.02	S08A	0.55	0.87
B03Z	0.43	0.99	S08B	0.59	0.83
B04Z	0.23	1.15	S08C	0.60	0.98
B05Z	0.35	1.08	S08D	0.47	1.08
B06Z	0.34	1.02	S08E	0.53	0.96
B07B	0.25	1.02	S08F	0.64	0.82
B07C	0.40	1.02	S08G	0.48	1.11
B08Z	0.42	1.14			
B09A	0.54	0.93			
B09B	0.63	0.93			
B09C	0.67	0.90			
B09D	0.51	0.92			
B09E	0.64	0.78			
B09F	0.61	0.81			
B09G	0.58	0.84			

2 That is, the correlation between the score on one item and the total raw score on all other items in the test.

Dimensionality and local dependence

After removing five items from the international scale, we used multidimensional item response models to assess the dimensionality of items. The two item dimensions that we explored corresponded to the structure of the cognitive domains (strands) described in the ICILS assessment framework (Fraillon et al., 2013).

To analyze test dimensionality, we first used the dimensions *collecting and managing information* (Strand 1) and *producing and exchanging information* (Strand 2), both of which are described in the ICILS assessment framework (see Fraillon et al., 2013). We also reviewed dimensionality by distinguishing items reflecting single tasks from items within large tasks and defining the four modules as separate dimensions.

The result of our multidimensional IRT modeling using ACER ConQuest showed latent correlations between the assessment framework dimensions of 0.96, thus indicating very high levels of similarity between the scales based on these different subgroups of items. Given these results, a decision was made not to report subscales reflecting assessment framework dimensions in the ICILS 2013 international report (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014).

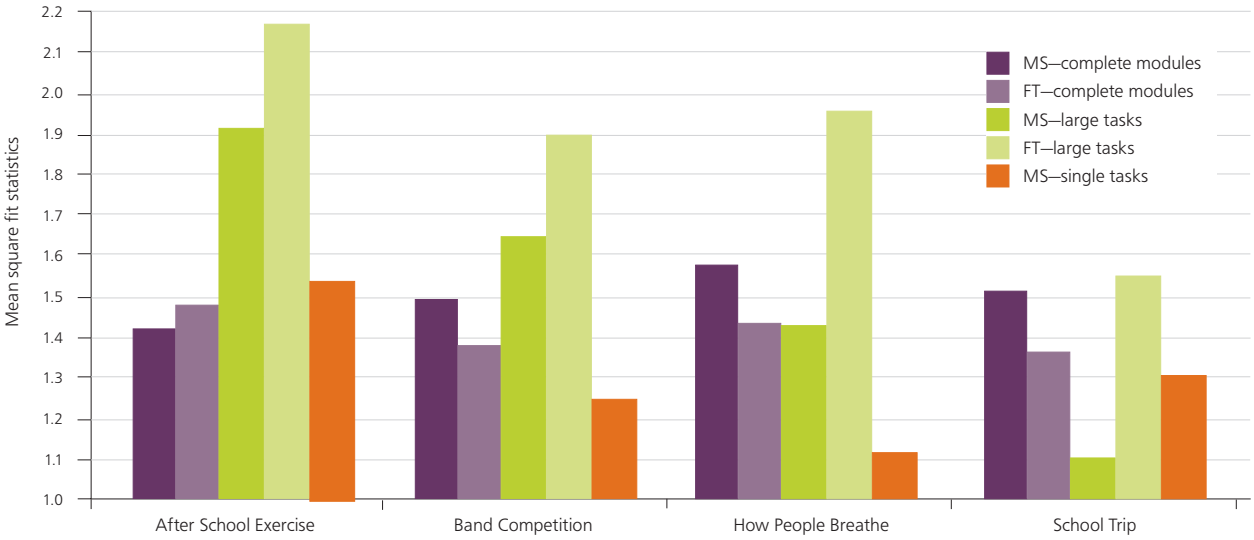
Estimates of latent correlations between other subsets of items, however, were somewhat lower. The latent correlation between performance on the single items and performance on items nested within large tasks was 0.76. The correlation between the After School Exercise task (module) and any of the other three modules was higher than 0.80. Correlations between the other three modules ranged from 0.75 to 0.77. These results suggest that the items within modules were more similar than the items from different modules to one another and that this outcome could be partly due to the nesting of items in large tasks as well as the common narrative for each module. In other words, the review of item dimensionality provided some evidence of local dependence between items within modules.

To investigate this matter further, we conducted an analysis of local dependence by estimating fit statistics for user-defined combinations of items (see Adams & Wu, 2009). These statistics are similar to the fit statistics estimated for each item parameter estimate. In this case, a fit statistic of 1.0 for a group of items suggests the (complete) absence of local dependence between the items in this group. A fit statistic of 1.3 or higher indicates local dependence.

Figure 11.5 presents the results from the ICILS main survey (darker shading) in comparison with those from the ICILS field trial (lighter shading). The fit statistics for complete modules were between 1.4 and 1.6. In all modules except School Trip, local dependence was highest for items nested within the large tasks. Comparison of the main survey to the field trial showed less local dependence within large tasks in the main survey, although the fit values above 1.3 for three out of four modules still indicated a certain level of local dependence.

We also undertook analyses designed to review local dependence between pairs of items within the single tasks or within 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 or to remove one of the two items from the scale. The review showed only one pair of items that had a weighted mean square statistic with a value above 1.3. We collapsed these two items, A10B and A10F in the After School Exercise module, into one item and assigned a score of 1.0 to students who responded correctly to both items.

Figure 11.5: Local dependence within modules



Assessment of scorer reliabilities

The scoring of open-ended items in the ICILS cognitive test was guided by scoring guides that were refined following the experiences in the international field trial. Within countries, subsamples of about 20 percent of student responses to each task were scored twice by 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 an assessment of scoring reliabilities. Table 11.2 shows the average percentages of scorer agreement across participating countries as recorded at this review stage. Percentages of agreement between double-scored items ranged from 72 to 98 percent across countries.

As has been the practice in other IEA studies, the only items included in the international database were those scored with scorer agreement of at least 70 percent.

Intercoder reliability (ICR) was also examined at the national level. Instances of an open-ended response item having scorer agreement below 70 percent (see Appendix D.1) were evident in 52 cases across 10 countries. The scores for the corresponding items were excluded during calibration of items and the drawing of plausible values but were included in the public database.

Differential item functioning by gender

This analysis included an exploration of the quality of the items. It involved assessing differential item functioning (DIF) by gender. DIF occurs when groups of students with the same degree of ability have different probabilities of responding correctly to an item. For example, if boys with the same degree of ability have a higher probability than girls of correctly answering an item, the item shows gender DIF. This situation is a violation of the model's assumptions, one of which is that the probability is a function of ability only and not of any group membership or other features.

Table 11.2: Percentages of scorer agreement for open-ended ICILS test items

Item	*N*	Agreement (%)	Item	*N*	Agreement (%)
S08G	5255	89	B09G	5524	85
S08F	5255	80	B09F	5524	97
S08E	5255	93	B09E	5524	96
S08D	5255	96	B09D	5524	92
S08C	5255	96	B09C	5524	83
S08B	5255	98	B09B	5524	86
S08A	5255	93	B09A	5524	92
H07J	5089	86	B02Z	5361	94
H07I	5089	89	B01Z	5467	91
H07H	5089	86	A10J	5278	93
H07G	5089	80	A10I	5278	85
H07E	5089	91	A10H	5278	80
H07D	5089	77	A10G	5278	74
H07C	5089	72	A10F	5278	94
H07B	5089	76	A10E	5278	88
H07A	5089	95	A10D	5278	91
H06Z	5537	97	A10C	5278	73
H05Z	5476	94	A10B	5278	84
			A10A	5278	85
			A07Z	5518	91
			A06C	4954	95
			A06B	5054	97
			A06A	5233	94

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

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

For the purpose of measuring parameter equivalence across the two gender groups, we included an interaction effect in the scaling model, where θ_n was the estimated ability of person n and δ_i was the estimated location of item i, with 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 then be modeled as

$$P_{x_i}(\theta) = \frac{\exp \sum_{j=0}^{x_i} (\theta_n - (\delta_i - \eta_g + \lambda_{ig} + \tau_{ij}))}{\sum_{h=0}^{m_i} \exp \sum_{j=0}^{h} (\theta_n - (\delta_i - \eta_g + \lambda_{ig} + \tau_{ij}))} \quad x_i = 0, 1, 2, \ldots, m_i$$

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.

Tests of statistical significance did not provide appropriate criteria for identifying DIF due to the mostly large but also varying sample sizes across participating countries.

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

We therefore, we used a DIF value of 0.3 (approximately about one third of a standard deviation) as a criterion for identification of gender DIF.

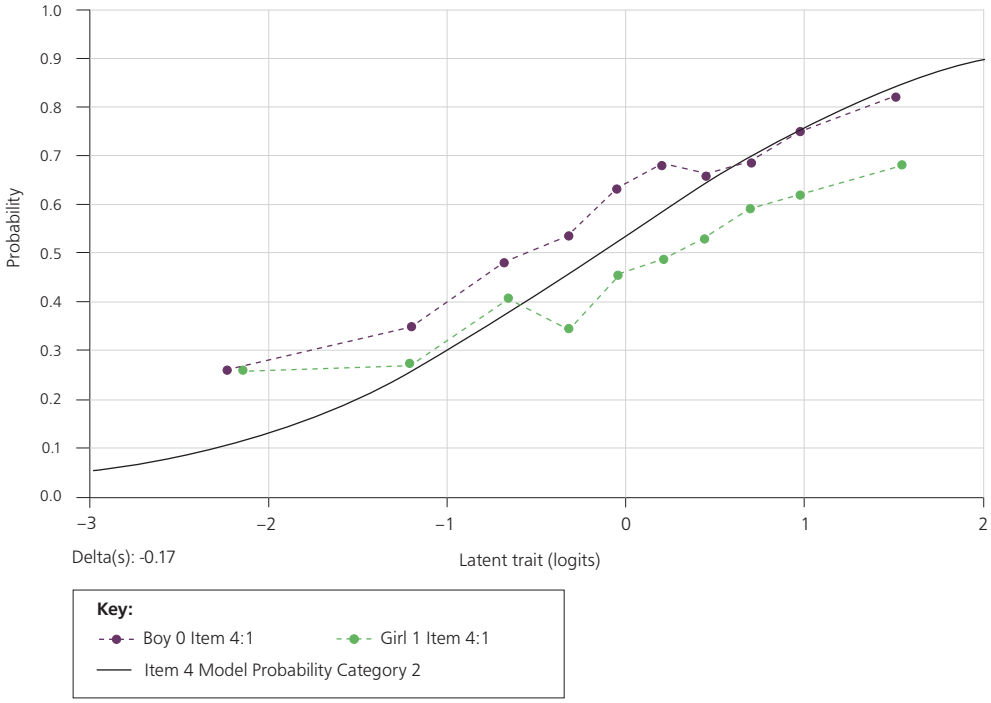
Table 11.3 shows the gender DIF estimates. Only one item showed noticeable DIF (with an estimate larger than 0.3 logits). This item (A04Z) was removed from the scale. Figure 11.6 illustrates the difference in the probability of males giving a correct response and females giving a correct response. Independent of their ability level, male students tended to have a higher probability than female students of responding correctly to this item.

Table 11.3: Gender DIF estimates for retained test items

Item	Female	Male	Difference (F-M)	Item	Female	Male	Difference (F-M)
A01Z	0.10	-0.10	0.20	H01Z	-0.09	0.09	-0.18
A02Z	-0.11	0.11	-0.22	H02Z	-0.06	0.06	-0.12
A03Z	0.14	-0.14	0.28	H03Z	-0.05	0.05	-0.10
A04Z	-0.32	0.32	-0.63	H05Z	0.00	0.00	-0.01
A05Z	-0.04	0.04	-0.08	H06Z	0.01	-0.01	0.02
A06A	-0.20	0.20	-0.40	H07A	0.04	-0.04	0.08
A06B	-0.01	0.01	-0.02	H07B	0.09	-0.09	0.18
A06C	-0.17	0.17	-0.34	H07C	0.04	-0.04	0.08
A07Z	-0.04	0.04	-0.08	H07D	-0.03	0.03	-0.05
A08Z	-0.09	0.09	-0.19	H07E	0.04	-0.04	0.07
A09Z	-0.10	0.10	-0.21	H07F	0.02	-0.02	0.03
A10A	0.15	-0.15	0.30	H07G	0.00	0.00	-0.01
A10B	-0.02	0.02	-0.03	H07H	0.02	-0.02	0.05
A10C	0.09	-0.09	0.18	H07I	0.04	-0.04	0.08
A10D	0.12	-0.12	0.25	H07J	-0.03	0.03	-0.07
A10E	0.10	-0.10	0.20	S01Z	-0.10	0.10	-0.20
A10F	-0.04	0.04	-0.07	S03Z	-0.02	0.02	-0.04
A10G	0.05	-0.05	0.09	S04A	-0.07	0.07	-0.13
A10H	0.07	-0.07	0.13	S04B	-0.01	0.01	-0.02
A10I	0.14	-0.14	0.29	S06Z	-0.05	0.05	-0.10
A10J	0.00	0.00	0.01	S07Z	-0.15	0.15	-0.30
B01Z	0.03	-0.03	0.06	S08A	0.10	-0.10	0.20
B02Z	-0.16	0.16	-0.32	S08B	0.14	-0.14	0.28
B03Z	-0.12	0.12	-0.23	S08C	0.03	-0.03	0.06
B04Z	-0.01	0.01	-0.02	S08D	-0.17	0.17	-0.34
B05Z	-0.01	0.01	-0.02	S08E	0.10	-0.10	0.19
B06Z	-0.02	0.02	-0.04	S08F	0.17	-0.17	0.33
B07B	0.15	-0.15	0.29				
B07C	0.05	-0.05	0.11				
B08Z	-0.19	0.19	-0.38				
B09A	0.04	-0.04	0.08				
B09B	0.05	-0.05	0.10				
B09C	0.01	-0.01	0.03				
B09D	0.12	-0.12	0.24				
B09E	0.06	-0.06	0.12				
B09F	0.06	-0.06	0.12				
B09G	0.06	-0.06	0.13				

Note: Items that were removed from the scale because of unsatisfactory scaling properties were not included in this analysis.

Figure 11.6: Gender DIF in Item A04Z



National reports with item statistics

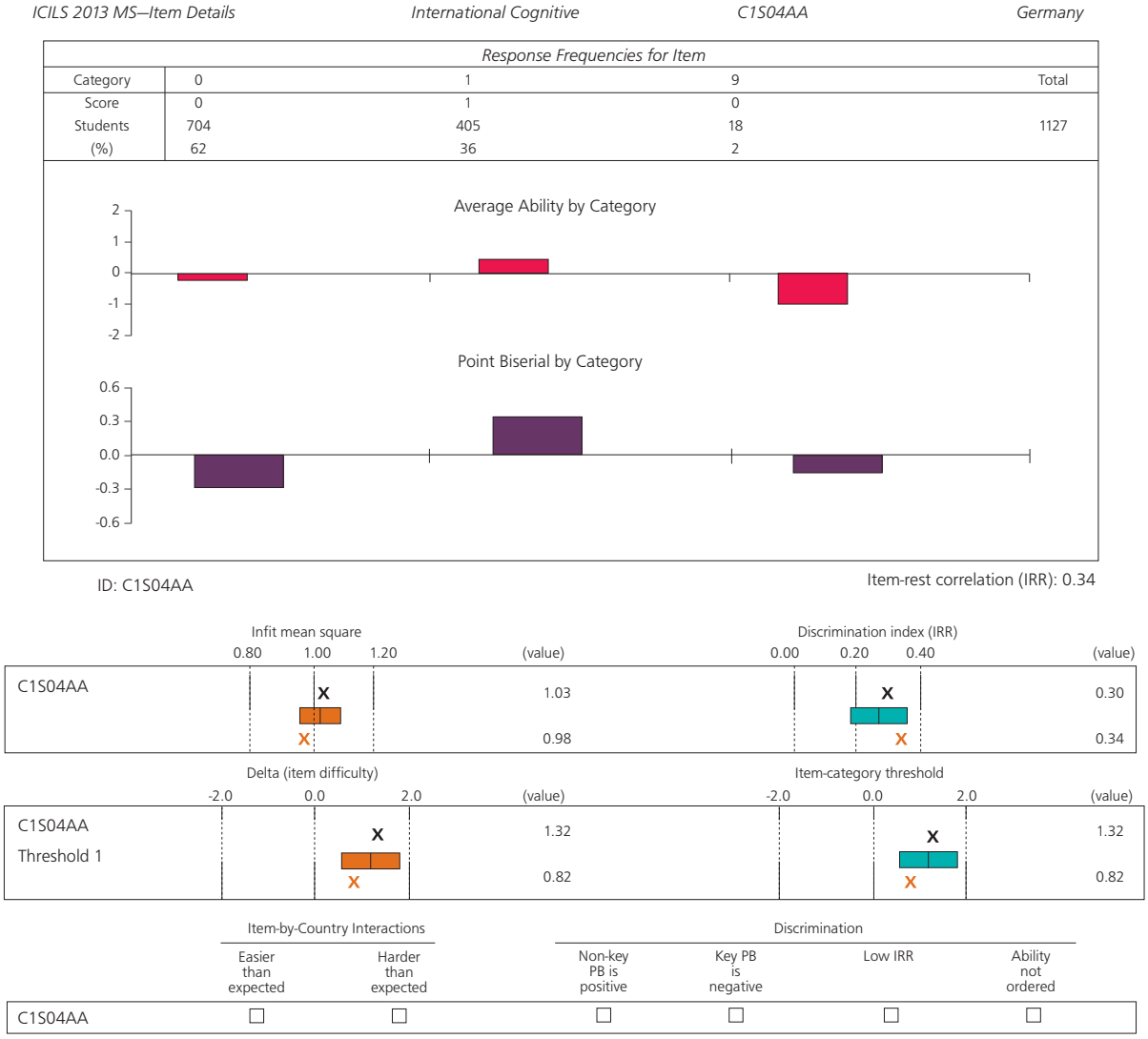
The international study center (ISC) at the Australian Council for Educational Research (ACER) provided national centers with item statistics (see the example in Figure 11.7) and asked them to review the test items flagged due to unsatisfactory scaling properties (e.g., negative correlations between correct response and overall score) or because of large differences between the national and international item difficulties. Items were also flagged whenever the category-total correlations were disordered. In some cases, national centers informed the ISC about translation errors (i.e., errors that had not been detected during verification), issues with scoring, and technical problems with computer-based delivery of particular items. We categorized the corresponding items as "not administered" in the international database (but only for the particular country) and excluded them from scaling of the country's national data (see Appendix D.1).

Working independently from those conducting the national item reviews, ISC staff reviewed national items that showed poor scaling properties (item misfit or large item-by-country interactions) by conducting post-verifications of item translation. In a number of cases, we identified additional national items that needed to be set to "not administered" in the international database and thus excluded from scaling of the corresponding national data. Appendix D.1 provides details about the items set to not administered in the database and omitted from scaling.

Crossnational measurement equivalence

With any test used to assess student achievement crossnationally, it is important that the test items function similarly across those countries. Items show *item-by-country interaction* when students from different countries but with the same ability vary in their probability of answering these questions. Test items with considerable item-by-country interaction are not suitable for scaling cognitive test items in international surveys.

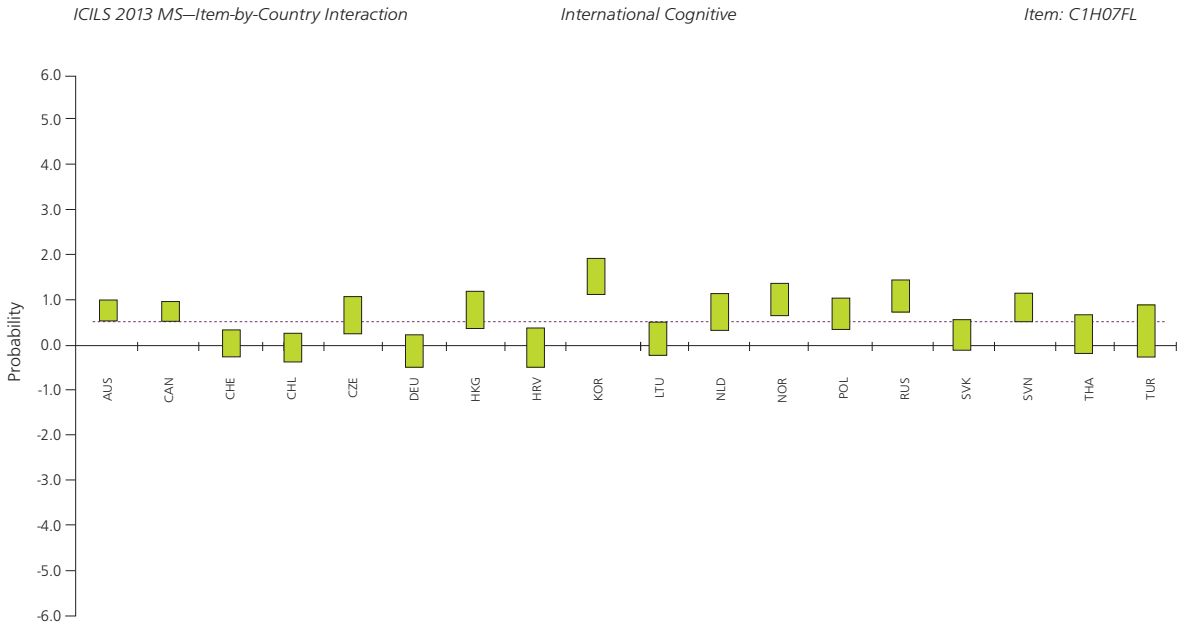
Figure 11.7: Example of item statistics provided to national centers



During our analyses of the ICILS main-survey test items, we compared national calibrations with international item parameters in order to assess the occurrence of item-by-country interaction. We also computed confidence intervals for each national item parameter, basing the computation on the respective standard errors and adjusting them for possible design effects and multiple comparisons.

As an example, Figure 11.8 shows the item-by-country interaction graph for Item H07F, which was the only item removed from the scale because of the large number of countries with statistically significant item-by-country DIF. The figure shows clear and considerable variation in item difficulties 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 crossnational DIF was used to identify items for postverification checks after completion of the main data collection.

Figure 11.8: Example of item-by-country interaction graph for Item H07F



Although the ICILS test items showed generally only limited item-by-country interactions, some national item difficulties deviated quite considerably (more than 1.3 logits) from the international item difficulty. In these cases, we omitted the items from scaling for those national samples where larger deviations had been observed. Appendix D.1 records the items that were omitted nationally from scaling due to high levels of item-by-country interaction.

Missing data issues

Initially, there were three possible types of missing responses in the ICILS test data. These were omitted items (coded as 9), not-administered items (coded as 8), and not-reached items (coded as 7). An item response was assigned Code 7 if a student did not respond to any of the items within the same booklet[4] that followed that item (i.e., the student did not complete any of the remaining test questions). This code was also assigned if the student did not respond to the item preceding it.

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

The omitted response category (Code 9) was used when a student provided no response at all to an item administered to him or her. Not-administered item responses (Code 8) were assigned to items that had been included in the item pool but not in the booklet administered to a student due to the rotation of modules, or (in a very few cases) where item responses were missing due to technical failure or incorrect translations of the item content.

For analysis purposes, the third missing category, "not reached," was divided into two categories at the post-processing stage:

- Response patterns that aligned with the not-reached pattern were temporarily assigned a code of 6 if a student had not run out of time or if test administrators recorded a technical failure. Here it was assumed that these students could not finish each module in the test because of a problem in the delivery system. These responses were treated as not administered in the scaling model.
- Not-reached responses retained Code 7 if the student had no time remaining to finish a module in the booklet. The extent of occurrence of Code 7 items provided us with information about the appropriateness of the test's length as well as the appropriateness of its difficulty.

Table 11.4 shows the international percentages of omitted, not-reached, and technical-failure responses.

Table 11.5 shows the average percentages of missing values overall, by item type, and for each module. On average, each item was omitted by 12 percent of the students. However, there was considerable variation between multiple-choice and constructed-response items, which were automatically scored, and large tasks, with the latter omitted more often by students than the other item types (14% and 13%, respectively).

The average number of students who did not reach an item was 0.0 percent and who could not respond to an item due to technical failure was 0.3 percent. Because all modules finished with a large task, the large tasks had the highest percentage of not-reached responses and technical failures.

The test module How People Breathe resulted in the highest percentage of omitted responses (16%) and Band Competition the lowest (6%). Average percentages of not-reached responses and items with technical failure were close to zero for most modules. Again, How People Breathe showed somewhat higher percentages than the other modules.

International item calibration and test reliability

Item parameters were obtained from calibration samples consisting of student subsamples from each country with a selection probability proportional to their sampling weight value. The calibration of item parameters involved randomly selecting subsamples of 500 students from each country that had met sample participation requirements. The two Canadian provinces were represented with a common subsample of 500 students. This process ensured that each country that had met sample participation requirements was equally represented in the sample. The final calibration sample included data from 7,500 students from 15 countries.

Table 11.4: Percentages of omitted responses and items not reached due to lack of time or technical failure for test items

Item	Omitted	Not Reached	Technical Failure	Item Type
A01Z	0.0	0.0	0.0	CMC
A02Z	36.2	0.0	0.0	CR auto
A03Z	2.9	0.0	0.0	MC
A04Z	2.4	0.0	0.0	MC
A05Z	16.5	0.0	0.0	CR auto
A06A	13.4	0.0	0.0	CR human
A06B	17.5	0.0	0.0	CR human
A06C	20.4	0.0	0.0	CR human
A07Z	5.1	0.0	0.0	CR human
A08Z	11.7	0.0	0.0	CR auto
A09Z	0.1	0.0	0.1	CR auto
A10A	14.9	0.0	0.1	Large task
A10B	14.9	0.0	0.1	Large task
A10C	14.9	0.0	0.1	Large task
A10D	14.9	0.0	0.1	Large task
A10E	14.9	0.0	0.1	Large task
A10F	14.9	0.0	0.1	Large task
A10G	14.9	0.0	0.1	Large task
A10H	14.9	0.0	0.1	Large task
A10I	14.9	0.0	0.1	Large task
A10J	14.9	0.0	0.1	Large task
A10K	14.9	0.0	0.1	CR auto
B01Z	8.0	0.0	0.0	CR human
B02Z	11.1	0.0	0.0	CR human
B03Z	45.4	0.0	0.0	CR auto
B04Z	14.5	0.0	0.0	MC
B05Z	4.5	0.0	0.0	CR auto
B06Z	16.6	0.0	0.0	CR auto
B07A	0.0	0.0	0.0	CR auto
B07B	0.0	0.0	0.0	CR auto
B07C	0.0	0.0	0.0	CR auto
B08Z	3.0	0.0	0.0	CMC
B09A	6.4	0.0	0.0	Large task
B09B	6.4	0.0	0.0	Large task
B09C	6.4	0.0	0.0	Large task
B09D	6.4	0.0	0.0	Large task
B09E	6.4	0.0	0.0	Large task
B09F	6.4	0.0	0.0	Large task
B09G	6.4	0.0	0.0	Large task
H01Z	22.5	0.0	0.0	CR auto
H02Z	0.0	0.0	0.0	CR auto
H03Z	6.3	0.0	0.0	CR auto
H04A	5.3	0.0	0.2	CR human
H05Z	4.0	0.1	0.6	CR human
H06Z	2.9	0.2	1.1	CR human
H07A	16.1	0.2	1.1	Large task
H07B	16.1	0.2	1.1	Large task
H07C	16.1	0.2	1.1	Large task
H07D	16.1	0.2	1.1	Large task
H07E	16.1	0.2	1.1	Large task
H07F	16.1	0.2	1.1	Large task
H07G	16.1	0.2	1.1	Large task
H07H	16.1	0.2	1.1	Large task
H07I	16.1	0.2	1.1	Large task
H07J	16.1	0.2	1.1	Large task
S01Z	17.6	0.0	0.0	CR auto
S02Z	22.9	0.0	0.0	CR auto
S03Z	40.1	0.0	0.0	CR auto
S04A	3.3	0.0	0.1	CR auto
S04B	3.3	0.0	0.1	CR auto
S05Z	5.5	0.0	0.1	MC
S06Z	2.3	0.1	0.3	CMC
S07Z	12.6	0.1	0.4	MC
S08A	10.0	0.1	0.4	Large task
S08B	10.0	0.1	0.4	Large task
S08C	10.0	0.1	0.4	Large task
S08D	10.0	0.1	0.4	Large task
S08E	10.0	0.1	0.4	Large task
S08F	10.0	0.1	0.4	Large task
S08G	10.0	0.1	0.4	Large task

Table 11.5: Percentages of omitted and invalid responses overall, by item type and by module

	Average Percent Correct		
	Omitted	Not reached	Technical failure
Grand average	**11.7**	**0.0**	**0.3**
Item type			
Multiple-choice	7.6	0.0	0.1
Complex multiple-choice	1.8	0.0	0.1
Constructed-response (human)	9.8	0.0	0.2
Constructed-response (auto)	13.8	0.0	0.0
Large task	12.5	0.1	0.4
Module			
After-School Exercise	14.9	0.0	0.1
Band Competition	6.4	0.0	0.0
How People Breathe	16.1	0.2	1.1
School Trip	10.0	0.1	0.4

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. However, they were included and treated as "incorrect" during scaling of the student responses. Technical failures were treated as not administered.

Data from countries that did not meet the sampling requirements after inclusion of replacement schools (i.e., Category 3 countries; see Chapter 7 for details) were not included in the calibration of item parameters. Table 11.6 shows the final item parameters based on the international calibration sample that we used to scale the ICILS test data. The table also records the standard errors for these parameters and for valid categories with scoring rules.

The overall reliability of the international test, as obtained from the scaling model, was 0.89 (ACER ConQuest 3.0 estimate).

Ability 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 booklets and respond to a set of items obtained from the main pool of items, has become standard in assessments of this type. 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 employing plausible value methodology that uses all available information from student tests and questionnaires, a process that leads to more accurate population as well as subgroup estimates (Mislevy, 1991; Mislevy & Sheehan, 1987; von Davier, Gonzalez, & Mislevy, 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

Table 11.6: Final parameter estimates of the international test items

Item #	Item Code	Difficulty (Logits)	Step Parameters	Difficulty (Scale Score)	Codes	Scoring/Key
1	A01Z	-0.91		475	0,1,2,3,4	0,0,0,0,1
2	A02Z	0.09		559	0,1	
3	A03Z	-2.70		324	1,2,3,4	0,0,1,0
4	A05Z	-1.91		390	0,1	
5	A06A	1.85		707	0,1	
6	A06B	1.12		645	0,1	
7	A06C	2.31		746	0,1	
8	A07Z	-0.18		536	0,1	
9	A08Z	-3.08		292	0,1	
10	A09Z	-0.23		532	0,1,2	0,0,1
11	A10A	-0.36	0.39	521	0,1,2	
12	A10BF	0.47		591	0,1,2,3	0,0,0,1
13	A10C	0.73	-0.43	613	0,1,2	
14	A10D	0.15	-0.96	564	0,1,2	
15	A10E	-1.58		418	0,1	
16	A10G	1.51	-0.06	679	0,1,2	
17	A10H	0.42	-0.16	586	0,1,2	
18	A10I	1.09		643	0,1	
19	A10J	0.14		563	0,1	
20	B01Z	-1.07		461	0,1	
21	B02Z	-0.55		505	0,1,2	0,0,1
22	B03Z	0.21		569	0,1	
23	B04Z	0.85		623	1,2,3,4	0,0,1,0
24	B05Z	0.06		556	0,1	
25	B06Z	-2.17		368	0,1	
26	B07B	-3.07		293	0,1	
27	B07C	-0.81		483	0,1	
28	B08Z	0.53	-1.00	596	0,1,2,3,4,5	0,0,0,0,1,2
29	B09A	0.46	-1.70	590	0,1,2	
30	B09B	-0.33	0.18	523	0,1,2	
31	B09C	-0.31	0.26	525	0,1,2	
32	B09D	0.11		561	0,1,2	0,1,1
33	B09E	-0.59		501	0,1	
34	B09F	-0.18		536	0,1,2	0,1,1
35	B09G	0.01		552	0,1	
36	H01Z	-0.48		511	0,1	
37	H02Z	0.98		634	0,1,2	0,0,1
38	H03Z	-1.35		437	0,1	
39	H05Z	1.71		696	0,1	
40	H06Z	0.96		632	0,1	
41	H07A	-1.51		424	0,1	
42	H07B	0.37	-0.15	582	0,1,2	
43	H07C	-0.13	-0.15	540	0,1,2	
44	H07D	-0.89	0.27	476	0,1,2	
45	H07E	0.31	-0.90	577	0,1,2	
46	H07G	1.05	0.11	640	0,1,2	
47	H07H	-0.12	0.58	541	0,1,2	
48	H07I	0.16		564	0,1	
49	H07J	2.56		767	0,1	
50	S01Z	-1.43		430	0,1	
51	S03Z	0.01		552	0,1	
52	S04A	1.30		661	0,1	
53	S04B	-0.61		500	0,1	
54	S06Z	-0.26		529	0,1,2	0,0,1
55	S07Z	2.59		770	A to K	G=1, Other=0
56	S08A	-0.44		514	0,1	
57	S08B	0.06		556	0,1	
58	S08C	0.10	2.93	559	0,1,2	
59	S08D	0.79	0.26	618	0,1,2,3	0,1,1,2
60	S08E	0.88	0.48	626	0,1,2	
61	S08F	1.08	-1.00	642	0,1,2	
62	S08G	0.26	-0.31	574	0,1,2	

item response model and the population model, which includes the regression on background variables used for conditioning. (For a detailed description, see Adams, Wu, & Macaskill, 1997; also Adams, 2002.) In order to obtain estimates of students' CIL, we used ACER Conquest 3.0 software, which allowed us to draw plausible values (see Wu et al., 2007).

We used all available international student questionnaire variables for conditioning of students. For missing responses, all missing values in a variable were substituted with either the mode or the mean, and extra variables were added to the additional indicators for missing values. Appendix D.2 lists all the international student-level variables (along with their respective scoring) that were used to condition the plausible values of CIL.

To reduce the large number of student-level variables, we included, as conditioning variables, principal components that reflected 99 percent of the variance in the original variables. At the student level, we included only gender and its missing indicator as direct conditioning variables. For between-school differences, we introduced the average of (weighted likelihood) ability estimates for all other students in the same school as well as other direct conditioning variables as indicator variables for explicit strata. This approach also allowed us to account for the two-level structure of the data.

After drawing plausible values, we transformed the resulting scale to a metric with a mean of 500 and a standard deviation of 100 for equally weighted ICILS countries that had met sampling requirements (Categories 1 and 2 countries), a process that meant excluding the benchmarking participants. This linear transformation was computed by applying the formula

$$\theta'_n = 500 + 100 \left(\frac{\theta_n - \bar{\theta}}{\sigma_\theta} \right)$$

where θ'_n were the student scores in the international metric, θ_n were the original logit scores, $\bar{\theta}$ was the international mean of student logit scores (-0.119) with equally weighted country subsamples, and σ_θ was its corresponding international standard deviation (1.186). We applied this transformation to each of the five plausible values. Chapter 13 provides a description of how plausible values were used to calculate imputation variance.

The development of proficiency levels for CIL

One of the objectives of ICILS was to establish a described CIL achievement scale that would become a reference point for future international assessments in this learning area. Establishing proficiency levels for 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.

Development of proficiency levels requires application of a method which ensures that the notion of "being at a level" can be interpreted consistently and in line with the fact that the achievement scale is a continuum. The ICILS research team therefore wanted not only to provide a common understanding about what being at a particular CIL level meant but also 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 level, which enabled us to create the properties described below:

- *The response probability (rp) for reporting item parameters*—set at $rp = 0.62$.
- *The width of the proficiency levels*—set at 0.8 logits.

Using these parameters, we were able to infer the following about students' aptitude in relation to the proficiency levels:

- Students whose results placed them at the lowest possible point of the proficiency level were likely to correctly answer (on average) slightly over 50 percent of the items on a test made up of items spread uniformly across the level, from the easiest to the most difficult item.
- Students whose results placed them at the lowest possible point of the 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 we chose 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 four proficiency levels that could be used when reporting student performances from the assessment. Table 11.7 shows the cut-points for these levels (in logits and final scale scores). The table also cites the percentage of students at each proficiency level across the participating ICILS countries.

Table 11.7: Proficiency level cut-points and percentage of students at each level

	Logits		Scale Scores		Percentages
	Higher than	Below or equal to	Higher than	Below or equal to	
Below Level 1		-1.7		407	17
Level 1	-1.7	-0.7	407	492	23
Level 2	-0.7	0.3	492	576	38
Level 3	0.3	1.3	576	661	21
Level 4	1.3		661		2

When reporting released test items and mapping them against proficiency levels, we had to transform the location parameters of these items to a value that reflected a response probability of 62 percent. We achieved this by adding the natural log of the odds of 62 percent chance to the original log odds and then transforming the result to the international metric by applying the same transformation as for the (original) student scores. The standardized item difficulty d'_i for each item that was obtained follows:

$$d'_i = 500 + 100 \times \left(\frac{d_i + 1n(0.62/0.38) - \bar{\theta}}{\sigma_\theta} \right)$$

Here, d_i is the item parameter in its original metric, $\bar{\theta}$ is the international mean of student logit scores, and σ_θ is its corresponding international standard deviation. These were used to standardize the plausible values. The standardized item parameters are included in Table 11.6.

Summary

The ICILS test items were scaled using item response modeling with the (one-parameter) Rasch model. Prior to scaling, we carried out an extensive analysis of scaling properties that included reviews of missing values, test coverage, assessment of item fit, differential item functioning by gender, and crossnational measurement equivalence.

We generated plausible values as ability estimates and conducted full conditioning in order to take all student-level and between-school differences into account. Four proficiency levels were established, thereby providing test item locations on the CIL achievement scale and allowing a description of these levels complete with example test items.

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. J., Wu, M. L., & 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* (pp. 111–145). Chestnut Hill, MA: Boston College.

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.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International Computer and Information Literacy Study assessment framework*. 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, NY: 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.

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.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). A*CER ConQuest: General item response modelling software* [computer program]. Camberwell, Victoria, Australia: Australian Council for Educational Research.

CHAPTER 12:

Scaling Procedures for ICILS Questionnaire Items

Wolfram Schulz and Tim Friedman

Introduction

This chapter describes the procedures used to scale the ICILS questionnaire data (for students, teachers, school principals, and ICT coordinators) and the indices based on them.

Two general types of indices could be distinguished, both of which derived from the ICILS questionnaires:

1. Simple indices constructed through arithmetical transformation or simple recoding, for example, ratios between ICT and students or an index of immigration background based on information about the country of birth of students and their parents; and
2. Scale indices derived from scaling of items, a process typically achieved by using item response modeling of dichotomous or Likert-type items.

The first part of this chapter describes the simple indices derived from the ICILS questionnaire data and the procedures applied to create them. The second part outlines the scaling procedures used in ICILS. The third and final part provides a detailed description of scaled indices, along with statistical information on item parameters, scale reliabilities, and the factor structure of related item sets.

The cross-country validity of item dimensionality and constructs was assessed during the field trial stage of ICILS. At this time, the international study center (ISC) at the Australian Council for Educational Research (ACER) used data to assess the extent to which measurement models held across participating countries. Similar to other IEA studies (see, for example, Schulz, 2009), the review made extensive use of both confirmatory factor analysis and item response modeling to examine crossnational measurement equivalence before the final selection of main survey questionnaire items was conducted.

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 the student tracking forms (see Chapter 8 for more details) provided the data needed to calculate 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., "2013" or "1998"), 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 their gender. Responses were recorded as the *sex of student* (S_SEX). Girls were coded as 1 and boys as 0. Gender information on the tracking form was imputed in those instances where students had omitted data for this question in their student questionnaire.

Question 3 asked students about their expected highest level of educational attainment. The corresponding index, *students' expected education* (S_ISCED), had the following categories:

1. No completion of ISCED 2;
2. Completion of ISCED 2 (lower-secondary);
3. Completion of ISCED Level 3 (upper-secondary);
4. Completion of ISCED 4 (nontertiary postsecondary) or ISCED 5B (vocational tertiary);
5. Completion of ISCED 5A (theoretically oriented tertiary) or ISCED 6 (postgraduate).

The ICILS student questionnaire collected information on the *country of birth* of the students and their parents (Question 4). For each student (S_SBORN) and his or her mother (S_MBORN) and father (S_FBORN), a code of 1 was given if they were born in the country of the assessment while a score of 0 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:

1. Students without immigrant background (students born in the country of assessment or who had at least one parent born in the country);[1]
2. Students born in the country of assessment but had both parents or only one parent born in another country;
3. Students born outside the country of assessment and with both parents or with only one parent born in another country.

We assigned missing values to students with missing responses for either their own place of birth, or that of their mother and father, or for all three questions. The analysis of immigrant background and CIL was based on a dichotomous indicator variable that distinguished between students with an immigrant background (Categories 2 and 3) and students without an immigrant background (Category 1). This variable was 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 the assessment or another language.[2] 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;
2. The language spoken at home most of the time differed from the language of the assessment.

1 Students who were born abroad but had at least one parent born in the country of the test were also classified as students without an immigrant background.

2 Most countries collected more detailed information on language use. This information is not included in the ICILS international database.

Occupational data for each student's two parents were obtained by first asking students whether their mother and father were in paid work or not (Questions 6 and 9), and then asking them to provide details as to what their parents' jobs were. Students used an open-response format to answer these questions (Questions 7 and 10).

The paid work status of the student's mother and father was classified into S_MWORK and S_FWORK respectively (1 indicating that the parent was in paid work and 0 indicating that the parent was not in paid work). The ICILS national centers coded the open-ended responses into four-digit codes using the International Standard Classification of Occupations (ISCO-2008) framework (International Labour Organization, 2007). These codes are contained in the indices S_MISCO (student's mother) and S_FISCO (student's father).

We then mapped these codes to the international socioeconomic index of occupational status (ISEI) (Ganzeboom, de Graaf, & Treiman, 1992). The three indices that we obtained from these scores were mother's occupational status (S_MISEI), father's occupational status (S_FISEI), 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 indicated higher levels of occupational status.

Questions 8 and 11 asked students to report on the *highest parental education attainment* of their mother and father respectively and so provided the data for measuring this important family background variable. The core difficulties with this variable related 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). In ICILS, we classified levels of parental education according to the International Standard Classification of Education (ISCED) (UNESCO, 2006).

Recoding educational qualifications into the categories below provided the following indices of highest parental educational attainment:

1. Did not complete ISCED 2;
2. ISCED 2 (lower-secondary);
3. ISCED 3 (upper-secondary);
4. ISCED 4 (nontertiary postsecondary) or ISCED 5B (vocational tertiary);
5. ISCED 5A (theoretically oriented tertiary) or ISCED 6 (postgraduate).

Indices with these categories are available for each student's mother (S_MISCED) and father (S_FISCED). The index for *highest educational level of parental education* (S_HISCED) corresponds to the higher ISCED level of either parent.

Question 12 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:

1. 0 to 10 books;
2. 11 to 25 books;
3. 26 to 100 books;
4. 101 to 200 books;
5. More than 200 books.

Teacher questionnaire

Teacher gender (T_SEX) was computed from the data captured from Question 1 of the teacher questionnaire. Female teaches were coded as 1; 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 "under 25" a value of 23 and coded "60 or over" as 63.

Question 5 of the teacher questionnaire asked teachers to indicate how long they had been *using computers for teaching purposes*. Responses to this question were used to form the basis of ICT experience in years of teaching (T_EXPT). This index was coded as:

1. Never;
2. Fewer than two years;
3. Two years or more.

School questionnaires

The first question of the ICILS school principal questionnaire asked respondents whether they were male or female. This information was used to form the index *gender of principal* (P_SEX); female principals were coded as 1, and male principals were coded as 0.

Question 3 of the principal questionnaire asked principals to record the number of girls and the number of boys in the entire school (IP1G03A, IP1G03B), while Question 4 asked them to record the number of girls and the number of boys enrolled in the target grade (IP1G04A, IP1G04B). 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 the number of students in the target grade (P_NUMTAR).

Question 5 also asked principals to report the lowest (youngest) (IP1G05A) grade and the highest (oldest) (IP1G05B) grade taught in their school. The difference between these two grades was calculated as the *number of grades in school* (P_NGRADE).

Question 6 collected information on the *number of teachers* (P_NUMTCH) in the school. The index was calculated by summing the total number of fulltime teachers (IP1G06A) with the total number of parttime teachers weighted at 50 percent (0.5 x IP1G06B). 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 8 collected information about whether the school was a public school or a private school. This information was used to form a *private school indicator* (P_PRIV), where public schools were coded as 0 and private schools were coded as 1.

Question 10 of the school principal questionnaire acted as a filter question for subsequent questions on the school's ICT *use for teaching and learning activities* (P_ICTLRN). Schools that used ICT for such activities were coded as 1 and asked to continue to the next question. Schools that did not use ICT for such activities were coded as 0 and asked to proceed to Question 14.

Question 3 of the school ICT coordinator questionnaire asked respondents to indicate the number of years ICT had been used in the school. Response categories for this question, *ICT experience in years in the school* (C_EXP), were as follows:

1. Never, we do not use computers;
2. Fewer than five years;
3. At least five but fewer than 10 years;
4. Ten years or more.

Question 7 of the ICT coordinator questionnaire collected data on the total number of computers in the school, the number of computers available to students, the number of computers with access to the internet/World Wide Web, and the number of smartboards or interactive whiteboards available at the school. In conjunction with the number of students at school (P_NUMSTD), these data provided the following ratios:

- *Ratio of school size and number of computers* (C_RATCOM) = number of students in the school (P_NUMSTD)/number of computers in the school altogether (IIG07A).
- *Ratio of school size and number of computers available for students* (C_RATSTD) = number of students in the school (P_NUMSTD)/number of computers in the school available to students (IIG07B).
- *Ratio of school size and number of computers with access to internet/World Wide Web* (C_RATWWW) = number of students in the school (P_NUMSTD)/number of computers in the school connected to the internet/World Wide Web (IIG07C).
- *Ratio of school size and smartboards* (C_RATSMB) = number of students in the school (P_NUMSTD)/number of smartboards or interactive white boards available (IIG08).

Scaling procedures

Classic scaling analysis

This section reports reliabilities both overall and for national samples. We used Cronbach's alpha coefficient as an estimate of the internal consistency of each scale (Cronbach, 1951), for which values above 0.7 are typically regarded as satisfactory and those above 0.8 indicate high reliability (see, for example, Nunnally & Bernstein, 1994, pp. 264–265). In addition to determining scale reliabilities, we reviewed 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 score based on all other items in a scale (adjusted item-total correlations).

Confirmatory factor analysis

Structural equation modeling (SEM) (Kaplan, 2009) allows the confirmation of theoretically expected dimensions and, at the field trial stage, respecification of the expected dimensional structure.[3] 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

$$(1)\quad x = \Lambda_x \xi + \delta$$

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.

3 In the initial stages of field trial analyses, we also employed exploratory factor analysis (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Tucker & MacCallum, 1997) to determine item dimensionality of larger item pools.

When conducting the confirmatory factor analyses for ICILS questionnaire data, we selected model-fit indices that provided measures of the extent to which a particular model with an assumed a-priori structure "fitted the data." For the ICILS 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 & Long, 1993).

Typically, RMSEA values over 0.10 suggest an unacceptable fit, those between 0.08 and 0.1 a mediocre model fit, and values of 0.05 and lower a close model fit (MacCallum, Browne, & Sugawara, 1996). As additional fit indices, CFI and NNFI are bound between 0 and 1. Values below 0.90 indicate a nonsatisfactory model fit whereas values greater than 0.95 suggest a close model fit (see Bentler & Bonnet, 1980; Hu & Bentler, 1999).

In addition to these fit indices, standardized factor loadings and the corresponding residual item variances provided further evidence of model fit for questionnaire data. Standardized factor loadings λ' can be interpreted in the same way as standardized regression coefficients if the indicator variable is regressed on the latent factor. The loadings also reflect the extent to which each indicator measures the underlying construct. Squared standardized factor loadings indicate how much of the variance in an indicator variable can be explained by the latent factor. The loadings are related to the (standardized) residual variance estimate δ' (these provide an estimate of the unexplained proportion of variance) as

$$\delta' = (1 - \lambda'^2) .$$

The use of multidimensional models also allows an assessment of the estimated correlation(s) between latent factors. These provide 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 are continuous. Therefore, the ICILS analysis team relied on robust weighted least squares estimation (WLSMV) (see Flora & Curran, 2004; Muthén, du Toit, & Spisic, 1997) to estimate the confirmatory factor models. The software package used for estimation was MPlus 7 (Muthén & 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 which met sample participation requirements (see Chapter 7 for further information). National samples of students, teachers, and schools received weights that ensured equal representations of countries in the analyses.

Item response modeling

Item response modeling provided an appropriate way of scaling 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

$$P_i(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \qquad (0)$$

where $P_i(\theta)$ 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 (k) categories (as, for example, with Likert-type items), this model can be generalized to the (Rasch) *partial credit model* (Masters & Wright, 1997), which takes the form of

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

where $P_{x_i}(\theta)$ denotes the probability of person n scoring x on item i, θ_n denotes the person's latent trait, the item parameter δ_i gives the location of the item on the latent continuum, and τ_{ij} denotes an additional step parameter.

Weighted mean-square statistics (*infit*), which are statistics based on model residuals, provided a way of assessing general fit to the scaling model. The residual-based statistics in conjunction with a wide range of further item statistics provided the basis for an assessment of item response theory IRT model fit. ICILS used the ACER Conquest software package (Wu, Adams, Wilson, & Haldane, 2007) to analyze the item-scaling properties and the estimation of item parameters.

The international item parameters came from the following calibration samples, which consisted of randomly selected subsamples or equally weighted national datasets:

A. *Calibration of student item parameters:* This was done using subsamples of 500 students randomly selected from each (weighted) national sample that met sample participation requirements for the student survey. The calibration sample included data for 14 countries[4] and from the two benchmarking participants (the Canadian provinces of Newfoundland and Labrador, and Ontario), which were represented with a (combined) subsample of 500 Canadian students. Subsamples were drawn to reduce the database for calibration.

B. *Calibration of teacher item parameters:* This was done using all data from 12 countries[5] and one benchmarking participant (Newfoundland and Labrador) that had met sample participation requirements for the teacher survey, thus giving each of the national samples an equal weight.

C. *Calibration of school item parameters:* This was done based on all data from the 14 countries and the combined Canadian sample, all of which had met sample participation requirements for the student survey, thus giving each of the national samples an equal weight.

After completing the estimation of the international item parameters from the calibration sample, we computed weighted likelihood estimations in order to obtain individual student scores. Weighted likelihood estimations are computed by minimizing the equation

4 These countries were Australia, Chile, Croatia, the Czech Republic, Germany, Republic of Korea, Lithuania, Norway, Poland, the Russian Federation, the Slovak Republic, Slovenia, Thailand, and Turkey.

5 These countries, except Germany and Norway, which did not meet sample participation requirements for the teacher survey, were the same as those in the calibration sample for the student questionnaire data.

$$\sum_{i\in\Omega}\left[\left(r_x+\frac{J_n}{2I_n}\right)-\sum_{j=1}^{k}\frac{\exp(\sum_{j=0}^{x}(\theta_n-\delta_i+\tau_{ij})}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{k}(\theta_n-\delta_i+\tau_{ij})}\right]=0 \qquad (0)$$

for each case n, where r_x is the sum score obtained from a set of k items with j 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). ACER ConQuest software allowed us to precalibrate item parameters in order to derive scale scores.

The transformation of weighted likelihood estimates to an international metric resulted in reporting scales with an ICILS average of 50 and a standard deviation of 10 for equally weighted datasets from the countries that met sample participation requirements. This transformation is achieved by applying the following formula:

$$\theta'_n = 50 + 10\,\frac{\theta_n-\bar{\theta}_{ICILS}}{\sigma_{\theta\,(ICILS)}}$$

where θ'_n are the scores in the international metric, θ_n are the original weighted likelihood estimates in logits, $\bar{\theta}_{ICILS}$ is the international mean of logit scores with equally weighted country subsamples, and $\sigma_{\theta(ICILS)}$ is the corresponding international standard deviation of the original weighted likelihood estimates. Appendix E of this report contains the means and standard deviations used to transform the original scale scores for the international student, teacher, and school questionnaires into the international metric.

Describing questionnaire scale indices

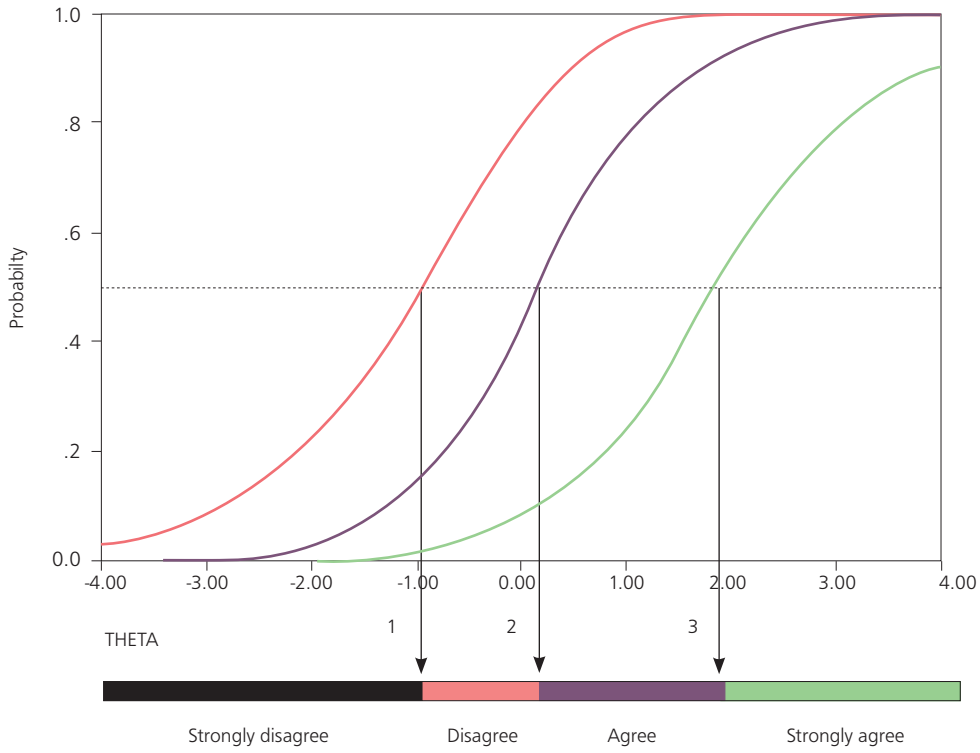
Questionnaire scales derived from weighted likelihood estimates (logits) presented values on a continuum with an ICILS average of 50 and a standard deviation of 10 (for equally weighted national samples). This presentation made it possible to interpret these scores by comparing individual scores or group average scores with the ICILS average. However, 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 allowed descriptions of these scales to be developed because scale scores could be mapped to (the expected) item responses.[6]

It is possible to describe item characteristics by using the parameters of the partial credit model in order to estimate the probability of each category being chosen relative to the probabilities of all higher categories being chosen. 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. The three vertical lines denote those points on the latent continuum where it becomes more likely to score > 0, > 1, or > 2. These locations τ_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.

6 This approach was also used in the IEA ICCS 2009 survey (see Schulz & Friedman, 2011).

Figure 12.1: Summed category probabilities for fictitious item



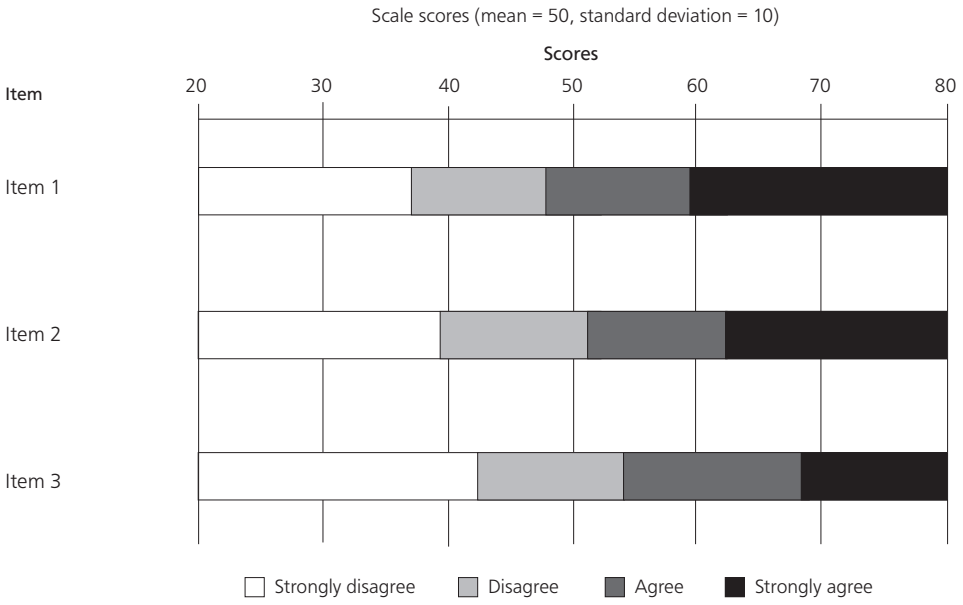
Summed probabilities are not identical to expected item scores and have to be understood in terms of the probability of scoring at least a particular category.[7] Thurstonian thresholds can be used to indicate, for each item category, those points on a scale at which respondents have a 0.5 probability of scoring this category or higher. For example, in the case of Likert-type items with the response categories *strongly disagree* (SD), *disagree* (D), *agree* (A), and *strongly agree* (SA), we can determine at what point of a scale a respondent has a 50 percent likelihood of agreeing with the item.

The item-by-score maps included in ICILS 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.

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 thus allows us to predict 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). The ICILS team used this approach with items measuring agreement. The approach allowed the team 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.

7 Other ways of describing item characteristics based on the partial credit model are *item characteristic curves*, which involve plotting the individual category probabilities, and *expected item score curves*. For a detailed description, see Masters and Wright (1997).

Figure 12.2: Example of questionnaire item-by-score map



Example of how to interpret the item-by-score map

#1: A respondent with a score of 30 has more than a 50% probability of strongly disagreeing with all three items

#2: A respondent with a score of 40 has more than a 50% probability of *not* strongly disagreeing with Items 1 and 2 but of strongly disagreeing 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 a score of 60 has more than a 50% probability of strongly agreeing with Item 1 and of at least agreeing with Items 2 and 3

#5: A respondent with a score of 70 has more than a 50% probability of strongly agreeing with Items 1, 2, and 3

In the reporting tables for questionnaire scales (depicted in the ICILS international reports), we depicted national average scale scores as boxes that indicated their mean values plus or minus sampling error and that were set in graphical displays featuring two underlying colors. National average scores located in the area set in (say) light blue on average across items would indicate that student responses had resided in the lower item categories ("disagree or strongly disagree," "not at all or not very interested," "never or rarely"). If these scores were found in the darker blue area, however, then we could assume students' average item responses would have been in the upper item response categories ("agree or strongly agree," "quite or very interested," "sometimes or often").

Scaled indices

Student questionnaire

National index of students' socioeconomic background

The multilevel analyses presented in the international report (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014) 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 produce an indicator variable with four categories: lower-secondary or below, upper-secondary, tertiary nonuniversity, and university education. We reduced the S_HOMLIT index from five to four categories (0 to 10 books, 11 to 25 books, 26 to 100 books, more than 100 books) by collapsing the two highest categories. We took this approach with the indices on parental education and home literacy because prior analyses showed approximately linear associations across these categories with CIL test scores and the other indicators of socioeconomic background.

In order to impute values for students who had missing data for only 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. We carried out this imputation procedure separately for each national sample.

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), we conducted principal component analysis of these indicator variables separately for each weighted national sample.

The final S_NISB scores consisted of factor scores for the first principal component with national averages of 0 and national standard deviations of 1. Table 12.1 shows the factor loadings and reliabilities for each national sample.

Students' use of ICT applications

The student questionnaire included three questions that required students to rate the frequency of their use of ICT applications. The following four scales were derived from these questions:

- Students' use of specific ICT applications (S_USEAPP);
- Students' use of ICT for social communication (S_USECOM);
- Students' use of ICT for exchanging information (S_USEINF);
- Students' use of ICT for recreation (S_USEREC).

Question 18 asked students to indicate the frequency with which they *used computer-based, work-oriented applications (software) outside school.* The response categories were "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." All seven items were used to derive the scale S_USEAPP, which had a reliability (Cronbach's alpha) of 0.80 across participating countries and coefficients ranging from 0.75 to 0.87. The higher values on the scale indicate more frequent use of computer applications.

Table 12.1: Factor loadings and reliabilities for national index of socioeconomic background

Country	Highest Parental Occupation	Highest Parental Education	Books at Home	Cronbach's Alpha
Australia	0.78	0.78	0.66	0.59
Chile	0.84	0.85	0.62	0.67
City of Buenos Aires, Argentina	0.85	0.85	0.67	0.71
Croatia	0.83	0.85	0.68	0.69
Czech Republic	0.82	0.80	0.63	0.62
Denmark	0.79	0.79	0.70	0.64
Germany	0.83	0.80	0.69	0.67
Hong Kong SAR	0.80	0.82	0.66	0.63
Korea, Republic of	0.73	0.76	0.64	0.51
Lithuania	0.79	0.82	0.72	0.67
Newfoundland and Labrador, Canada	0.75	0.81	0.59	0.54
Norway (Grade 9)	0.77	0.78	0.70	0.61
Ontario, Canada	0.78	0.79	0.61	0.56
Poland	0.83	0.85	0.67	0.69
Russian Federation	0.81	0.80	0.59	0.59
Slovak Republic	0.83	0.82	0.69	0.68
Slovenia	0.84	0.82	0.69	0.69
Switzerland	0.81	0.79	0.71	0.66
Thailand	0.83	0.83	0.61	0.63
Turkey	0.78	0.83	0.71	0.67
Average	**0.80**	**0.81**	**0.66**	**0.63**

Note: Benchmarking participants in italics.

Question 19 asked students to identify the frequency with which they were using the internet for a variety of communication and information-exchange activities outside of school. The response categories were "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." Exploratory factor analyses indicated that a two-dimensional model existed resulting in two scales reflecting *students' use of the internet for social communication* (S_USECOM) and for *exchanging information* (S_USEINF), each of them based on four items; two items did not load on any of the two dimensions. S_USECOM had an average reliability of 0.74, with the range extending from 0.67 to 0.81, while S_USEINF had an internal consistency of 0.73, with the range spanning 0.57 to 0.83. The higher values on both scales reflect more frequent use of the internet.

Question 20 required students to use the following response options to indicate how often they used computers for specified recreational purposes: "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." Five of the six items were used to derive a scale reflecting *students' use of ICT for recreation* (S_USEREC). This scale had an average reliability of 0.75, with the range extending from 0.66 to 0.85 across the participating countries and benchmarking participants. The higher scale values indicate more frequent use of ICT for recreation.

Figure 12.3 illustrates the results of the confirmatory factor analysis, which assumed a four-dimensional model with items from all four scales. The model fit was satisfactory, and moderate to high correlations were found between the four latent factors. Table 12.2 shows the scale reliabilities (Cronbach's alpha) for all four scales reflecting students' (out-of-school) use of ICT applications. These reliabilities were deemed satisfactory for most ICILS countries. Table 12.3 shows the item parameters for each of the four scales that were used to derive the IRT scale scores.

Figure 12.3: Confirmatory factor analysis of items measuring students' use of ICT applications

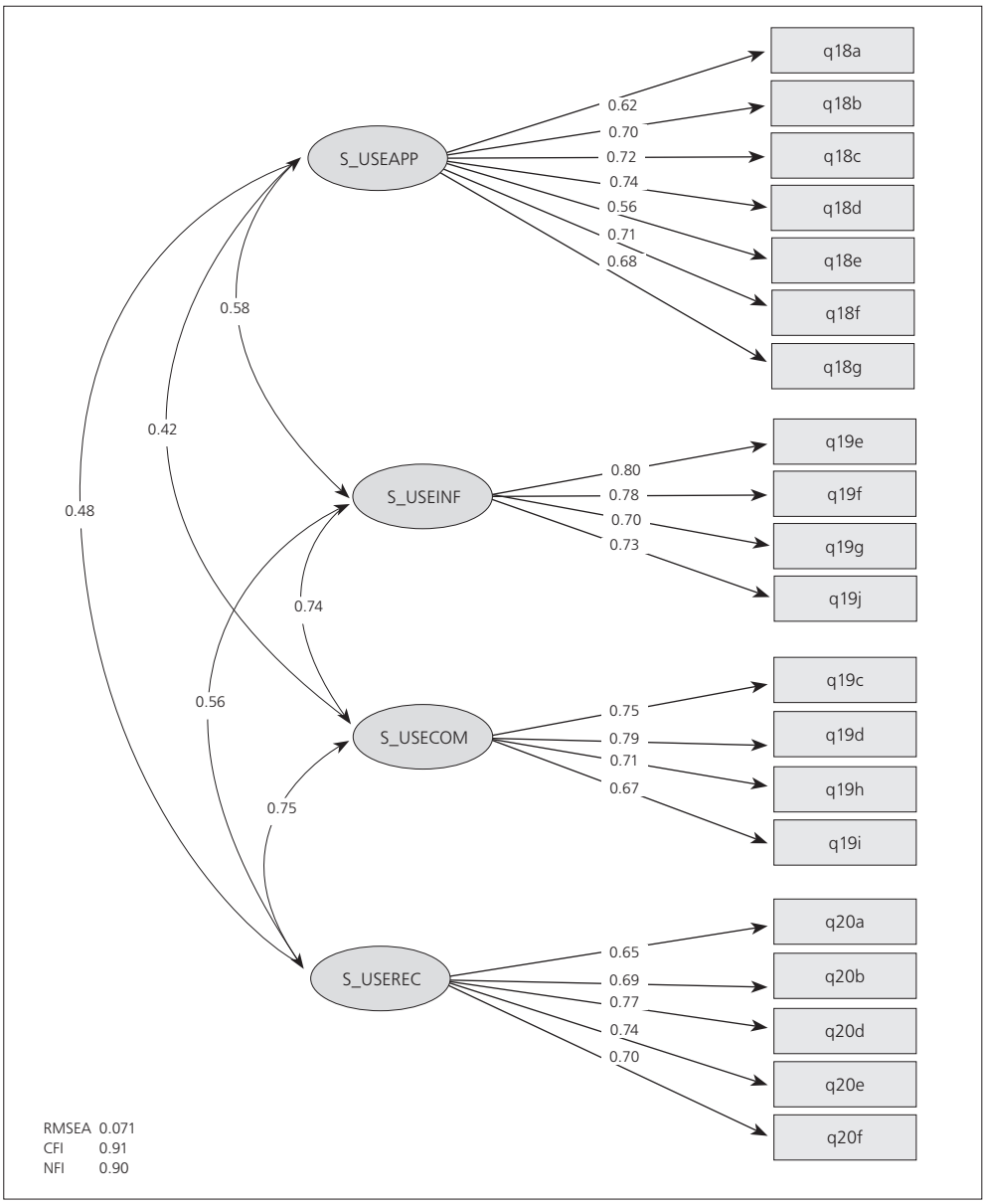


Table 12.2: Reliabilities for scales measuring students' use of ICT applications

Country	S_USEAPP	S_USECOM	S_USEINF	S_USEREC
Australia	0.80	0.74	0.77	0.78
Chile	0.80	0.73	0.75	0.76
City of Buenos Aires, Argentina	0.80	0.73	0.69	0.72
Croatia	0.80	0.75	0.73	0.76
Czech Republic	0.76	0.74	0.71	0.70
Denmark	0.75	0.71	0.57	0.71
Germany	0.79	0.72	0.70	0.66
Hong Kong SAR	0.87	0.76	0.72	0.82
Korea, Republic of	0.87	0.80	0.77	0.82
Lithuania	0.77	0.73	0.70	0.81
Netherlands	0.78	0.73	0.67	0.71
Newfoundland and Labrador, Canada	0.79	0.71	0.78	0.76
Norway (Grade 9)	0.80	0.71	0.64	0.70
Ontario, Canada	0.82	0.74	0.79	0.77
Poland	0.75	0.67	0.71	0.75
Russian Federation	0.80	0.75	0.71	0.76
Slovak Republic	0.79	0.74	0.71	0.75
Slovenia	0.81	0.79	0.78	0.74
Switzerland	0.77	0.71	0.77	0.69
Thailand	0.82	0.80	0.79	0.75
Turkey	0.84	0.81	0.83	0.85
Average reliability	**0.80**	**0.74**	**0.73**	**0.75**

Note: Benchmarking participants in italics.

Students' school-related ICT use

The student questionnaire included three questions measuring students' school-related use of ICT. The following three scales were derived from these questions:

- Students' use of ICT for (school-related) study purposes (S_USESTD);
- Students' use of ICT during lessons at school (S_USELRN);
- Students' reports on learning ICT tasks at school (S_TSKLRN).

Question 21 asked students to report how often they used computers for specified school-related purposes ("never," "less than once a month," "at least once a month but not every week," and "at least once a week"). All eight items were used to derive a scale reflecting s*tudents' use of ICT for (school-related) study purposes.* The scale had an average reliability of 0.83, and the range was from 0.78 to 0.91 across national samples. The higher values on the scale reflect more frequent use of ICT for study purposes.

Question 22 inquired how often students used computers during lessons in designated subjects or subject areas. The five response options for this question were "never," "in some lessons," "in most lessons," "in every or almost every lesson," and "I don't study this subject/these subjects." Student responses in the last category were treated as missing responses. Five of the eight items (referring to lessons in the subject areas most studied across all participating countries) were used to derive a scale reflecting *students' use of ICT during lessons* (S_USELRN). This scale had an average reliability of 0.81, and the coefficients ranged from 0.71 to 0.92 across national samples. The higher values on the scale reflect more frequent use of ICT during lessons at school.

Table 12.3: Item parameters for scales measuring students' use of ICT applications

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)	Tau(4)
S_USEAPP	*How often do you use a computer outside of school for each of the following activities?*					
IS1G18A	Creating or editing documents	-0.51	-1.54	-0.70	0.15	2.08
IS1G18B	Using a spreadsheet to do calculations, store data, or plot graphs	0.32	-1.25	-0.28	0.08	1.45
IS1G18C	Creating a simple "slideshow" presentation (for example, using [Microsoft PowerPoint ®])	-0.09	-1.91	-0.70	0.48	2.13
IS1G18D	Creating a multimedia presentation (with sound, pictures, video)	-0.01	-1.23	-0.15	0.20	1.19
IS1G18E	Using education software that is designed to help with your school study	0.01	-0.66	-0.34	-0.27	1.27
IS1G18F	Writing computer programs, macros, or scripts (for example, using [Logo, Basic, or HTML])	0.41	-0.47	-0.30	-0.20	0.96
IS1G18G	Using drawing, painting, or graphics software	-0.13	-1.14	-0.14	0.02	1.27
S_USECOM	*How often do you use the internet outside of school for each of the following activities?*					
IS1G19C	Communicating with others using messaging or social networks	-0.68	0.08	0.19	-0.34	0.08
IS1G19D	Posting comments to online profiles or blogs	0.15	-0.08	-0.15	-0.43	0.66
IS1G19H	Uploading images or video to an [online profile] or [online community]	0.35	-0.68	-0.18	-0.07	0.93
IS1G19I	Using voice chat (for example, Skype) to chat with friends or family online	0.18	-0.06	-0.18	-0.27	0.51
S_USEINF	*How often do you use the internet outside of school for each of the following activities?*					
IS1G19E	Asking questions on forums or [question and answer] websites	-0.13	-0.20	-0.25	-0.33	0.78
IS1G19F	Answering other people's questions on forums or websites	-0.16	-0.01	-0.14	-0.38	0.54
IS1G19G	Writing posts for your own blog	-0.05	0.37	-0.27	-0.57	0.47
IS1G19J	Building or editing a webpage	0.34	-0.12	-0.15	-0.19	0.45
S_USEREC	*How often do you use the internet outside of school for each of the following activities?*					
IS1G20A	Accessing the internet to find out about places to go or activities to do	0.72	-1.38	-0.27	0.21	1.44
IS1G20B	Reading reviews on the internet of things you might want to buy	0.75	-0.93	-0.52	0.03	1.42
IS1G20D	Listening to music	-0.86	-0.08	0.08	-0.28	0.29
IS1G20E	Watching downloaded or streamed video (for example, movies, TV shows, or clips)	-0.34	-0.32	-0.26	-0.35	0.94
IS1G20F	Using the internet to get news about things I am interested in	-0.26	-0.70	-0.21	-0.03	0.94

Question 23 asked students whether they had learned ("yes," "no") how to do eight different ICT-related tasks at school. All items were used to derive a scale reflecting *students' learning of ICT tasks at school*. Here, the average reliability was 0.81, and the coefficients ranged from 0.70 to 0.91. The higher scale values reflect a greater incidence of learning ICT tasks at school.

Figure 12.4 shows the results from a confirmatory factor analysis of all scaled items measuring students' school-related use of ICT. The model fit was good, and we found moderate positive correlations between the three latent factors. As Table 12.4 shows, all three scales had satisfactory reliabilities across all national samples. Table 12.5 records the item parameters for scaling.

Figure 12.4: Confirmatory factor analysis of items measuring students' reports on ICT use for study and learning

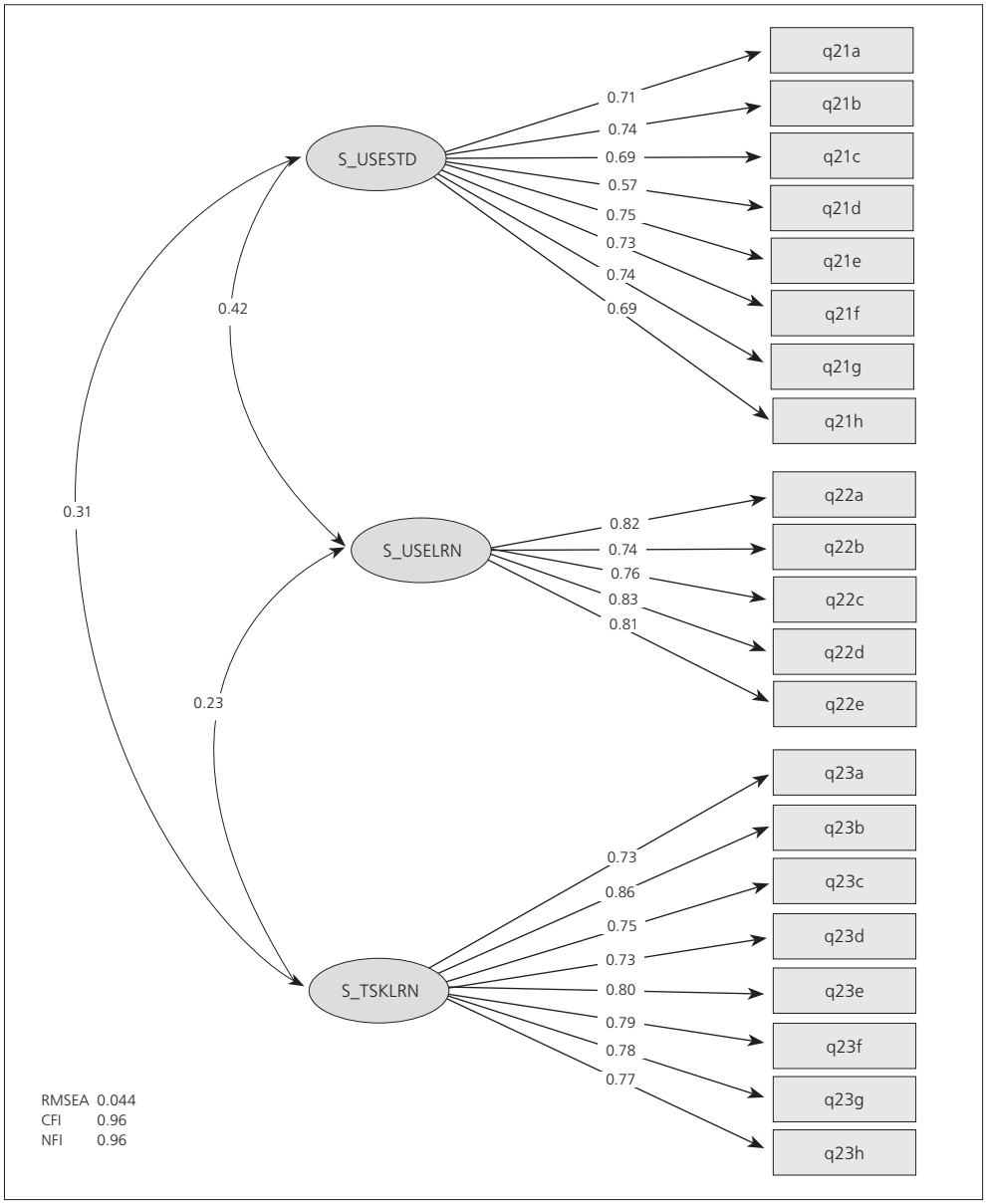


Table 12.4: Reliabilities for scales measuring students' reports on school-related use of ICT

Country	S_USESTD	S_USELRN	S_TSKLRN
Australia	0.84	0.79	0.70
Buenos Aires, Argentina	0.83	0.88	0.82
Chile	0.81	0.81	0.85
Croatia	0.81	0.84	0.84
Czech Republic	0.83	0.76	0.85
Denmark	0.78	0.84	0.78
Germany	0.80	0.71	0.76
Hong Kong SAR	0.89	0.92	0.87
Korea, Republic of	0.91	0.91	0.91
Lithuania	0.84	0.82	0.78
Netherlands	0.83	0.75	0.80
Newfoundland and Labrador, Canada	0.85	0.79	0.81
Norway (Grade 9)	0.81	0.71	0.82
Ontario, Canada	0.83	0.76	0.77
Poland	0.78	0.84	0.86
Russian Federation	0.82	0.88	0.80
Slovak Republic	0.80	0.79	0.76
Slovenia	0.83	0.79	0.83
Switzerland	0.82	0.72	0.80
Thailand	0.83	0.83	0.71
Turkey	0.88	0.85	0.80
Average reliability	**0.83**	**0.81**	**0.81**

Note: Benchmarking participants in italics.

Students' ICT self-efficacy, interest, and enjoyment

The student questionnaire included three questions measuring students' perceptions of their familiarity with and use of ICT. The following three scales were derived from these questions:

- Students' confidence (ICT self-efficacy) in solving basic computer-related tasks (S_BASEFF);
- Students' confidence (ICT self-efficacy) in solving advanced computer-related tasks (S_ADVEFF);
- Students' interest and enjoyment in using computers and computing (S_INTRST).

Question 25 asked students how well they thought they could do each of the 13 computer-based tasks. The response categories were "I know how to do this," "I could work out how to do this," and "I do not think I could do this." Two scales were derived from this item set. The first one reflected *students' self-confidence in solving basic computer-related tasks* (S_BASEFF), was based on six items, and had an average reliability of 0.76 across the national samples, with Cronbach's alpha coefficients ranging from 0.64 to 0.86. The second scale reflected *students' self-confidence in solving advanced computer-related tasks* (S_ADVEFF). It was derived from seven items, and the reliabilities ranged from 0.75 to 0.84, with an average of 0.80. The higher values on each scale reflect higher levels of ICT self-efficacy.

Table 12.5: Item parameters for scales measuring students' school-related use of ICT

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)
S_USESTD	*How often do you use computers for the following school-related purposes?*				
IS1G21A	Preparing reports or essays	-0.55	-1.43	0.06	1.37
IS1G21B	Preparing presentations	-0.59	-1.92	0.11	1.80
IS1G21C	Working with other students from your own school	-0.37	-1.41	0.05	1.36
IS1G21D	Working with other students from other schools	1.04	-0.44	-0.16	0.60
IS1G21E	Completing [worksheets] or exercises	-0.30	-1.08	0.05	1.03
IS1G21F	Organizing your time and work	0.10	-0.54	-0.05	0.60
IS1G21G	Writing about your learning	0.66	-0.43	-0.15	0.57
IS1G21H	Completing tests	0.01	-0.97	0.07	0.90
S_USELRN	*At school, how often do you use computers during lessons in the following subjects or subject areas?*				
IS1G22A	[Language arts: test language]	0.02	-1.67	0.86	0.81
IS1G22B	[Language arts: foreign and other national languages]	0.04	-1.75	0.85	0.90
IS1G22C	Mathematics	0.32	-1.41	0.72	0.70
IS1G22D	Sciences (general science and/or physics, chemistry, biology, geology, earth sciences)	-0.20	-1.62	0.64	0.98
IS1G22E	Human sciences/Humanities (history, geography, civics, law, economics, etc.)	-0.18	-1.62	0.63	0.98
S_TSKLRN	*At school, have you learned how to do the following tasks?*				
IS1G23A	Providing references to internet sources	0.06			
IS1G23B	Accessing information with a computer	-1.11			
IS1G23C	Presenting information for a given audience or purpose with a computer	-0.24			
IS1G23D	Working out whether to trust information from the internet	0.41			
IS1G23E	Deciding what information is relevant to include in school work	-0.02			
IS1G23F	Organizing information obtained from internet sources	0.13			
IS1G23G	Deciding where to look for information about an unfamiliar topic	0.23			
IS1G23H	Looking for different types of digital information on a topic	0.53			

Question 26 asked students to rate their agreement ("strongly agree," "agree," "disagree," "strongly disagree") with 11 statements. Seven of these items were used to derive a scale reflecting *students' interest in and enjoyment of computers and computing* (S_INTRST). The scale had an average reliability of 0.81 across the national samples, and the coefficients ranged from 0.74 to 0.86. The higher scores on the scale indicate higher levels of interest and enjoyment.

Figure 12.5 depicts the results from the confirmatory factor analysis of the scaled items from the two questions. There was a good fit for the three-factor model, and we found moderate to high positive correlations between the three latent factors. Table 12.6 shows the reliabilities for the three scales. These were satisfactory for most national samples. Table 12.7 records the item parameters used to scale these items.

Figure 12.5: Confirmatory factor analysis of items measuring students' ICT self-efficacy and interest/enjoyment

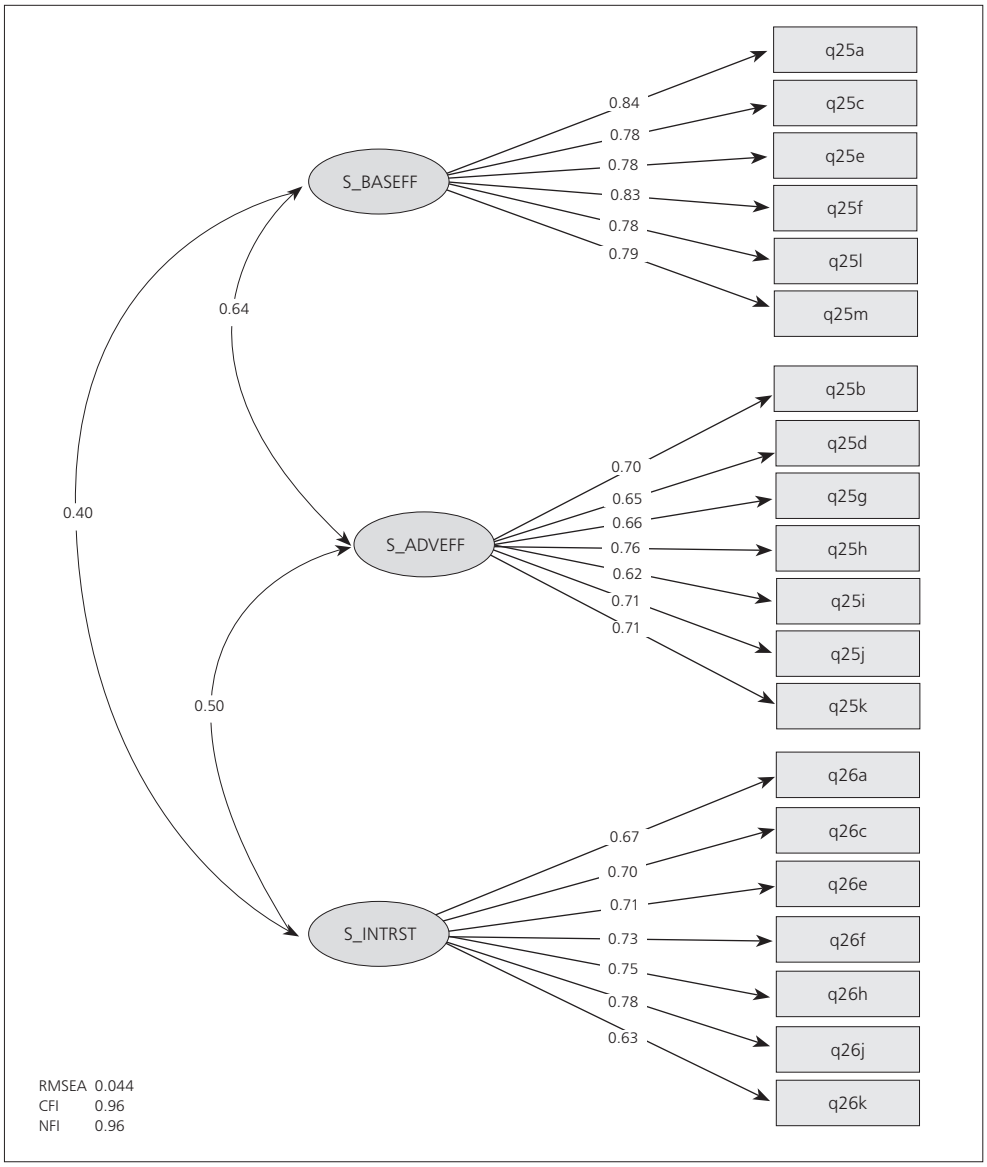


Table 12.6: Reliabilities for scales measuring students' ICT self-efficacy and interest/enjoyment

Country	S_BASEFF	S_ADVEFF	S_INTRST
Australia	0.73	0.79	0.85
Buenos Aires, Argentina	0.74	0.79	0.74
Chile	0.75	0.79	0.81
Croatia	0.78	0.79	0.81
Czech Republic	0.67	0.80	0.79
Denmark	0.64	0.79	0.82
Germany	0.80	0.81	0.80
Hong Kong SAR	0.86	0.82	0.86
Korea, Republic of	0.84	0.84	0.86
Lithuania	0.73	0.77	0.83
Netherlands	0.79	0.82	0.79
Newfoundland and Labrador, Canada	0.75	0.84	0.86
Norway (Grade 9)	0.69	0.83	0.83
Ontario, Canada	0.78	0.83	0.85
Poland	0.69	0.75	0.79
Russian Federation	0.81	0.76	0.74
Slovak Republic	0.72	0.80	0.78
Slovenia	0.78	0.80	0.81
Switzerland	0.78	0.81	0.85
Thailand	0.76	0.78	0.81
Turkey	0.84	0.83	0.84
Average reliability	**0.76**	**0.80**	**0.81**

Note: Benchmarking participants in italics.

Table 12.7: Item parameters for scales measuring students' ICT self-efficacy and interest/enjoyment

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)
S_BASEFF	*How well can you do each of these tasks on a computer?*				
IS1G25A	Search for and find a file on your computer	-0.70	-0.61	0.61	
IS1G25C	Edit digital photographs or other graphic images	0.35	-0.92	0.92	
IS1G25E	Create or edit documents (for example, assignments for school)	-0.08	-0.67	0.67	
IS1G25F	Search for and find information you need on the internet	-0.67	-0.39	0.39	
IS1G25L	Create a multimedia presentation (with sound, pictures, or video)	0.83	-1.06	1.06	
IS1G25M	Upload text, images, or video to an online profile	0.26	-0.59	0.59	
S_ADVEFF	*How well can you do each of these tasks on a computer?*				
IS1G25B	Use software to find and get rid of viruses	-0.18	-0.55	0.55	
IS1G25D	Create a database (for example, using [Microsoft Access ®])	0.48	-0.87	0.87	
IS1G25G	Build or edit a webpage	0.00	-0.91	0.91	
IS1G25H	Change the settings on your computer to improve the way it operates or to fix problems	-0.77	-0.69	0.69	
IS1G25I	Use a spreadsheet to do calculations, store data, or plot a graph	-0.82	-0.96	0.96	
IS1G25J	Create a computer program or macro (for example, in [Basic, Visual Basic])	0.95	-0.95	0.95	
IS1G25K	Set up a computer network	0.34	-0.56	0.56	
S_INTRST	*Thinking about your experience with computers: To what extent do you agree or disagree with the following statements?*				
IS1G26A	It is very important to me to work with a computer	-0.25	-1.74	-0.51	2.26
IS1G26C	I think using a computer is fun	-0.51	-1.50	-0.40	1.91
IS1G26E	It is more fun to do my work using a computer than without a computer	0.16	-1.59	-0.12	1.71
IS1G26F	I use a computer because I am very interested in the technology	0.97	-1.92	0.32	1.59
IS1G26H	I like learning how to do new things using a computer	-0.35	-1.42	-0.60	2.03
IS1G26J	I often look for new ways to do things using a computer	0.33	-2.08	0.08	2.00
IS1G26K	I enjoy using the internet to find out information	-0.35	-0.95	-0.83	1.78

Teacher questionnaire

Teachers' confidence in computer tasks (self-efficacy)

Question 7 of the ICILS teacher questionnaire required teachers to rate their ability to perform a series of tasks on a computer by themselves. This question was used to derive a scale reflecting *teachers' ICT self-efficacy* (T_EFF).

The response categories for teachers were "I know how to do this," "I could work out how to do this," and "I do not think I could do this." All 14 items in the question were used to derive scale T_EFF, which had a reliability (Cronbach's alpha) of 0.87 across participating countries and coefficients ranging from 0.82 to 0.94. The higher values on the scale indicate a greater degree of ICT self-efficacy.

Figure 12.6 illustrates the results of the confirmatory factor analysis assuming a one-dimensional model with items from the scale. The analysis showed a satisfactory model fit. As evident in Table 12.8, the scale reliabilities (Cronbach's alpha) for teachers' ICT self-efficacy were satisfactory across all national samples. Table 12.9 shows the item parameters for the scales that were used to derive the IRT scale scores.

Teachers' use of ICT applications for teaching

Question 9 of the ICILS teacher questionnaire asked teachers to indicate how often they used different ICT tools when teaching a given reference class. We used the items in this question to derive a scale (T_USEAPP) reflecting *teachers' use of ICT applications* in their reference class. The response categories for teachers were "never," "in some lessons," "in most lessons," and "in every or almost every lesson." The reliability (Cronbach's alpha) of the resulting scale (T_USEAPP) was 0.89 across participating countries, and the coefficients ranged from 0.81 to 0.96. The higher scale values indicate higher frequencies of teacher use of ICT tools in their reference class.

Figure 12.6: Confirmatory factor analysis of items measuring teachers' ICT self-efficacy

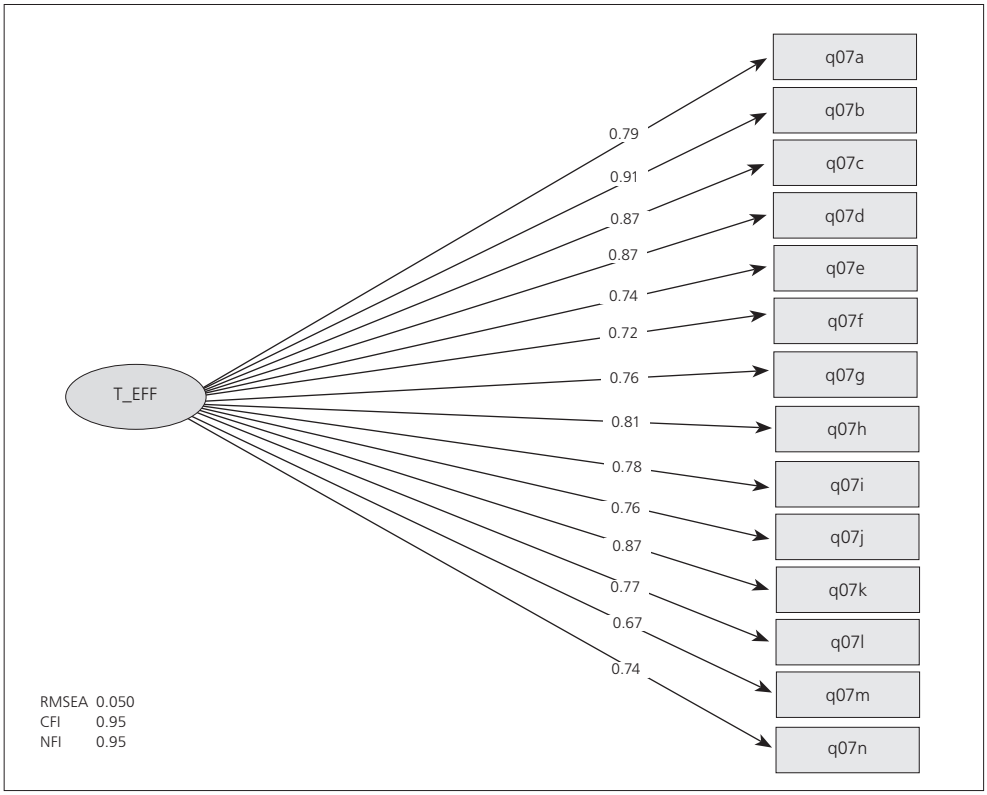


Table 12.8: Reliabilities for scale measuring teachers' confidence in computer tasks (self-efficacy)

Country	T_EFF
Australia	0.86
Chile	0.89
Croatia	0.94
Czech Republic	0.87
Denmark	0.82
Germany	0.87
Hong Kong SAR	0.86
Korea, Republic of	0.86
Lithuania	0.88
Netherlands	0.85
Newfoundland and Labrador, Canada	0.85
Norway (Grade 9)	0.83
Ontario, Canada	0.85
Poland	0.86
Russian Federation	0.92
Slovak Republic	0.90
Slovenia	0.87
Thailand	0.94
Turkey	0.91
Average reliability	**0.87**

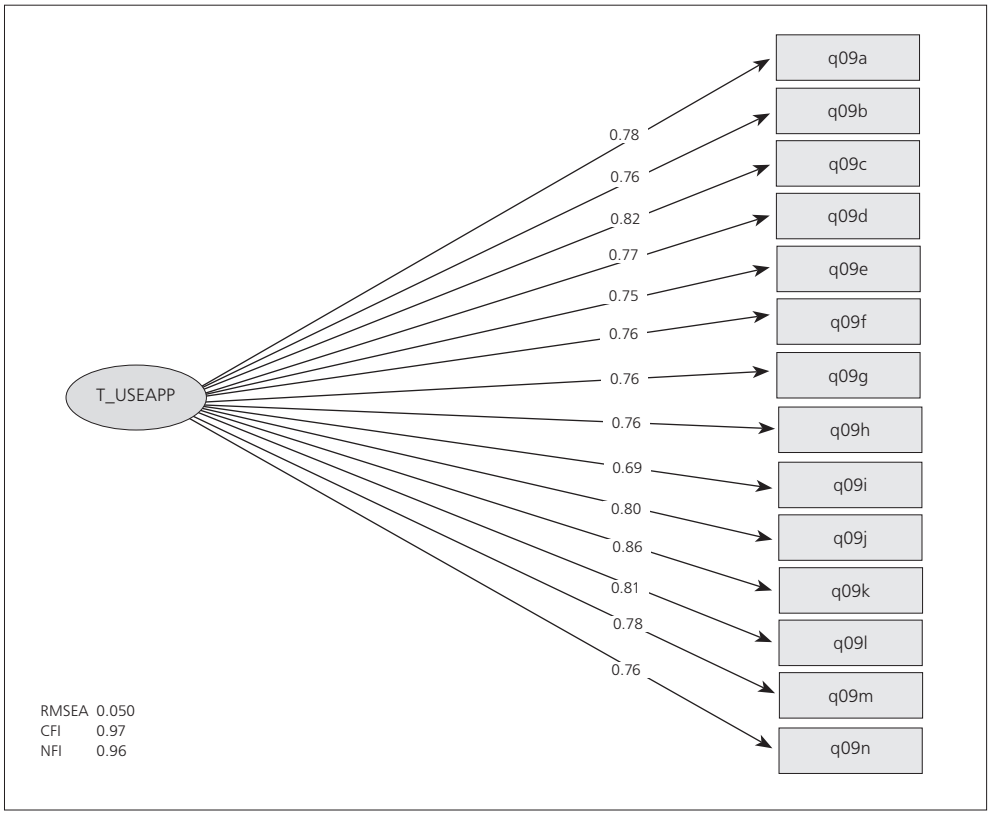
Note: Benchmarking participants in italics.

Table 12.9: Item parameters for scale measuring teachers' confidence in computer tasks (self-efficacy)

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)
T_EFF	*How well can you do these tasks on a computer by yourself?*			
IT1G07A	Searching for and finding a file on your computer	-1.10	-0.45	0.45
IT1G07B	Emailing a file as an attachment	-1.06	-0.30	0.30
IT1G07C	Storing your digital photos on a computer	-0.33	-0.42	0.42
IT1G07D	Filing digital documents in folders and subfolders	-0.44	-0.36	0.36
IT1G07E	Monitoring students' progress	0.15	-1.37	1.37
IT1G07F	Using a spreadsheet program (e.g. [Lotus 1 2 3 ®, Microsoft Excel ®]) for keeping records or analysing data	0.87	-1.12	1.12
IT1G07G	Contributing to a discussion forum/user group on the internet (for example, a wiki or blog)	0.76	-1.22	1.22
IT1G07H	Producing presentations (for example, [PowerPoint® or a similar program]), with simple animation functions	0.09	-0.61	0.61
IT1G07I	Using the internet for online purchases and payments	-0.18	-0.75	0.75
IT1G07J	Preparing lessons that involve the use of ICT by students	-0.05	-1.07	1.07
IT1G07K	Finding useful teaching resources on the internet	-1.66	-0.46	0.46
IT1G07L	Assessing student learning	-0.13	-1.27	1.27
IT1G07M	Collaborating with others using shared resources such as [Google Docs®]	1.32	-1.52	1.52
IT1G07N	Installing software	1.76	-0.70	0.70

Figure 12.7 illustrates the results of the confirmatory factor analysis assuming a one-dimensional model with the 14 items from the scale. The model fit was found to be satisfactory. Table 12.10 shows the scale reliabilities (Cronbach's alpha), which ranged from 0.81 to 0.96, with an average of 0.89. Table 12.11 shows the item parameters for the scale that was used to derive the IRT scale scores.

Figure 12.7: Confirmatory factor analysis of items measuring teachers' use of ICT applications in reference class



Teachers' use of ICT for activities and practices in class

The teacher questionnaire contained two questions about teachers' use of ICT in their reference class for a range of activities and practices. The two scales derived from these two questions were:

- Teachers' use of ICT for learning at school (T_USELRN);
- Teachers' use of ICT for teaching at school (T_USETCH).

Question 10 of the teacher questionnaire asked teachers to indicate how often they used 13 different activities in their reference class. The response options included "never," "sometimes," and "often." All 13 items were used to derive the scale titled *teachers' use of ICT for learning at school* (T_USELRN). The scale had an average reliability of 0.90 across national samples; the Cronbach's alpha coefficients ranged from 0.81 to 0.96. The higher scores on the scale indicate greater frequencies of use of such activities.

Question 11 of the teacher questionnaire asked teachers to indicate how often they used ICT for 11 different practices in their reference class. The response options were "never," "sometimes," and "often." Ten of the 11 items from this question (Items b to k) were used to derive the scale teachers' use of ICT for teaching at school (T_USETCH).

Table 12.10: Reliabilities for scale measuring teachers' use of ICT applications in reference class

Country	T_USEAPP
Australia	0.85
Chile	0.91
Croatia	0.91
Czech Republic	0.87
Denmark	0.86
Germany	0.87
Hong Kong SAR	0.89
Korea, Republic of	0.91
Lithuania	0.92
Netherlands	0.82
Newfoundland and Labrador, Canada	0.89
Norway (Grade 9)	0.81
Ontario, Canada	0.93
Poland	0.91
Russian Federation	0.93
Slovak Republic	0.92
Slovenia	0.89
Thailand	0.96
Turkey	0.95
Average reliability	**0.89**

Note: Benchmarking participants in italics.

Table 12.11: Item parameters for scale measuring teachers' use of ICT applications in reference class

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)
T_USEAPP	*How often did you use the following tools in your teaching of the reference class this school year?*				
IT1G09A	Tutorial software or [practice programs]	-0.80	-2.28	0.96	1.31
IT1G09B	Digital learning games	0.33	-2.48	1.09	1.39
IT1G09C	Wordprocessors or presentation software (for example, [Microsoft Word ®], [Microsoft PowerPoint ®])	-1.81	-2.17	0.65	1.53
IT1G09D	Spreadsheets (for example, [Microsoft Excel®])	0.28	-1.76	0.77	0.99
IT1G09E	Multimedia production tools (for example, media capture and editing, web production)	0.04	-1.59	0.60	1.00
IT1G09F	Concept mapping software (for example, [Inspiration ®], [Webspiration ®])	1.05	-1.30	0.35	0.96
IT1G09G	Data logging and monitoring tools	0.33	-1.32	0.68	0.63
IT1G09H	Simulations and modeling software	1.07	-1.57	0.51	1.05
IT1G09I	Social media (for example, Facebook, Twitter)	0.84	-1.20	0.58	0.62
IT1G09J	Communication software (for example, email, blogs)	-0.29	-2.09	0.85	1.23
IT1G09K	Computer-based information resources (for example, websites, wikis, encyclopedia)	-1.37	-2.62	0.71	1.91
IT1G09L	Interactive digital learning resources (for example, learning objects)	-0.65	-2.06	0.67	1.39
IT1G09M	Graphing or drawing software	0.21	-1.65	0.63	1.02
IT1G09N	e-portfolios	0.77	-1.18	0.50	0.68

The scale had a reliability (Cronbach's alpha) of 0.92, with the coefficients ranging from 0.86 to 0.97 across national samples. The higher values for this scale indicate more frequent use of the listed practices.

Figure 12.8 shows the results from a confirmatory factor analysis of all scaled items measuring teachers' use of ICT for activities and practices in their reference class. The model fit was very good, and we found a high positive correlation between the two latent factors (0.92). As Table 12.12 shows, both scales had satisfactory reliabilities across all national samples, with averages of 0.90 and 0.92 respectively. Table 12.13 records the item parameters used for scaling.

Figure 12.8: Confirmatory factor analysis of items measuring teachers' use of ICT for activities/ practices in reference class

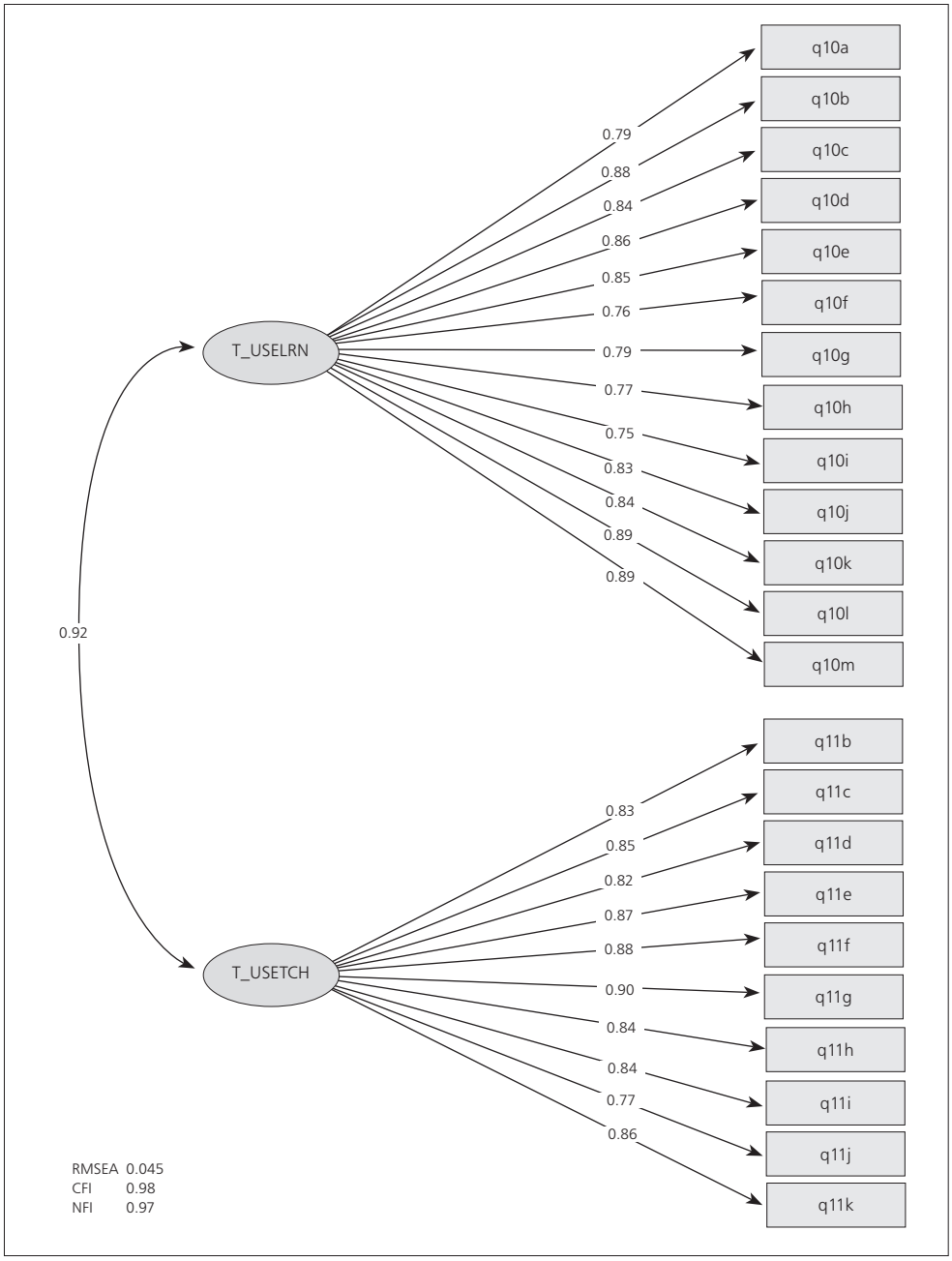


Table 12.12: Reliabilities for scales measuring teachers' use of ICT for activities and practices in class

Country	T_USELRN	T_USETCH
Australia	0.90	0.89
Chile	0.91	0.93
Croatia	0.91	0.93
Czech Republic	0.87	0.90
Denmark	0.86	0.89
Germany	0.87	0.90
Hong Kong SAR	0.89	0.92
Korea, Republic of	0.91	0.93
Lithuania	0.92	0.93
Netherlands	0.82	0.86
Newfoundland and Labrador, Canada	0.89	0.91
Norway (Grade 9)	0.81	0.89
Ontario, Canada	0.93	0.93
Poland	0.91	0.95
Russian Federation	0.93	0.94
Slovak Republic	0.92	0.93
Slovenia	0.89	0.90
Thailand	0.96	0.97
Turkey	0.95	0.97
Average reliability	**0.90**	**0.92**

Note: Benchmarking participants in italics.

Table 12.13: Item parameters for scales measuring teachers' use of ICT for activities and practices in class

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)
T_USELRN	*How often does your reference class use ICT in the following activities?*			
IT1G10A	Working on extended projects (i.e., over several weeks)	-0.23	-1.72	1.72
IT1G10B	Working on short assignments (i.e., within one week)	-1.31	-2.01	2.01
IT1G10C	Explaining and discussing ideas with other students	-0.08	-1.72	1.72
IT1G10D	Submitting completed work for assessment	-0.71	-1.64	1.64
IT1G10E	Working individually on learning materials at their own pace	-0.59	-1.78	1.78
IT1G10F	Undertaking open-ended investigations or field work	0.60	-1.51	1.51
IT1G10G	Reflecting on their learning experiences (for example, by using a learning log)	1.25	-1.13	1.13
IT1G10H	Communicating with students in other schools on projects	1.92	-1.22	1.22
IT1G10I	Seeking information from experts outside the school	0.73	-1.65	1.65
IT1G10J	Planning a sequence of learning activities for themselves	0.46	-1.33	1.33
IT1G10K	Processing and analyzing data	0.09	-1.71	1.71
IT1G10L	Searching for information on a topic using outside resources	-1.71	-1.68	1.68
IT1G10M	Evaluating information resulting from a search	-0.42	-1.76	1.76
T_USETCH	*How often does your reference class use ICT in the following activities?*			
IT1G11B	Providing remedial or enrichment support to individual students or small groups of students	-0.24	-1.75	1.75
IT1G11C	Enabling student-led whole-class discussions and presentations	-0.30	-1.84	1.84
IT1G11D	Assessing students' learning through tests	0.10	-1.34	1.34
IT1G11E	Providing feedback to students	-0.25	-1.58	1.58
IT1G11F	Reinforcing learning of skills through repetition of examples	-0.80	-1.71	1.71
IT1G11G	Supporting collaboration among students	-0.16	-1.62	1.62
IT1G11H	Mediating communication between students and experts or external mentors	1.94	-1.48	1.48
IT1G11I	Enabling students to collaborate with other students (within or outside school)	1.10	-1.57	1.57
IT1G11J	Collaborating with parents or guardians in supporting students' learning	0.62	-1.39	1.39
IT1G11K	Supporting inquiry learning	-0.10	-1.71	1.71

Teachers' emphasis on ICT in teaching

Question 12 of the ICILS teacher questionnaire required teachers to rate how much emphasis they gave to developing ICT-based capabilities in students in their reference class (the response categories were "strong emphasis," "some emphasis," "little emphasis," and "no emphasis"). All 12 items in the question were used to derive a scale reflecting teachers' emphasis on teaching ICT skills (T_EMPH).

The average reliability (Cronbach's alpha) of the scale was 0.97 across the participating countries, with coefficients ranging from 0.94 to 0.99. The higher scale scores indicate higher levels of teachers' emphasis on teaching ICT skills.

Figure 12.9 illustrates the results of the confirmatory factor analysis assuming a one-dimensional model with items from the scale. The model fit was satisfactory. Table 12.14 shows the scale reliabilities (Cronbach's alpha), and Table 12.15 shows the item parameters that were used to derive the IRT scale scores.

Figure 12.9: Confirmatory factor analysis of items measuring teachers' emphasis on ICT in teaching

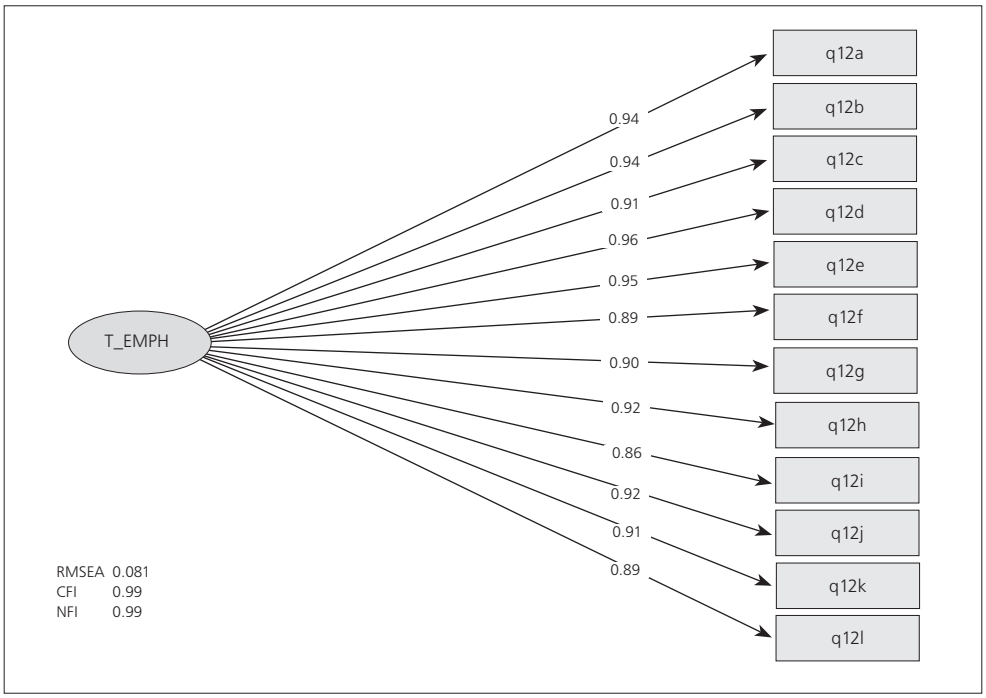


Table 12.14: Reliabilities for scale measuring teachers' emphasis on ICT in teaching

Country	T_EMPH
Australia	0.96
Chile	0.97
Croatia	0.99
Czech Republic	0.97
Denmark	0.96
Germany	0.96
Hong Kong SAR	0.97
Korea, Republic of	0.98
Lithuania	0.96
Netherlands	0.94
Newfoundland and Labrador, Canada	0.97
Norway (Grade 9)	0.95
Ontario, Canada	0.98
Poland	0.98
Russian Federation	0.97
Slovak Republic	0.98
Slovenia	0.96
Thailand	0.99
Turkey	0.99
Average reliability	**0.97**

Note: Benchmarking participants in italics.

Table 12.15: Item parameters for scale measuring teachers' emphasis on ICT in teaching

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)
T_EMPH	*In your teaching of the reference class in this school year how much emphasis have you given to developing the following ICT-based capabilities in your students?*				
IT1G12A	Accessing information efficiently	-1.09	-2.04	-0.76	2.80
IT1G12B	Evaluating the relevance of digital information	-0.11	-2.30	-0.48	2.79
IT1G12C	Displaying information for a given audience/purpose	-0.26	-2.27	-0.51	2.78
IT1G12D	Evaluating the credibility of digital information	-0.11	-2.05	-0.42	2.48
IT1G12E	Validating the accuracy of digital information	-0.01	-2.06	-0.40	2.46
IT1G12F	Sharing digital information with others	0.59	-2.36	-0.43	2.78
IT1G12G	Using computer software to construct digital work products (for example, presentations, documents, images, and diagrams)	-0.51	-1.90	-0.43	2.33
IT1G12H	Evaluating their approach to information searches	0.26	-2.29	-0.38	2.67
IT1G12I	Providing digital feedback on the work of others (such as classmates)	1.46	-2.36	-0.35	2.71
IT1G12J	Exploring a range of digital resources when searching for information	-0.14	-2.23	-0.49	2.72
IT1G12K	Providing references for digital information sources	0.12	-2.23	-0.34	2.57
IT1G12L	Understanding the consequences of making information publically available online	-0.19	-1.79	-0.19	1.98

Teachers' views on using ICT for teaching and learning

Question 13 of the ICILS teacher questionnaire asked respondents to indicate the extent to which they agreed or disagreed with statements about using ICT in teaching and learning (response categories were "strongly agree," "agree," "disagree," and "strongly disagree"). The items from this question provided data for deriving two scales:

- Positive views on using ICT in teaching and learning (T_VWPOS);
- Negative views on using ICT in teaching and learning (T_VWNEG).

The first scale (T_VWPOS) was derived from eight of the 15 items in the question and had an average reliability (Cronbach's alpha) of 0.83, with the coefficients ranging from 0.74 to 0.88 across participating countries. The remaining seven items were used to derive the second scale (T_VWNEG). It had an average reliability of 0.80, and the coefficients ranged from 0.72 to 0.87. The higher scores on these two scales represent stronger positive views and stronger negative views respectively.

Figure 12.10 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, and we found a moderate negative correlation between the two latent factors. Table 12.16 shows the scale reliabilities (Cronbach's alpha) for the scales reflecting teachers' views of ICT for teaching and learning at school. The reliabilities were satisfactory in most ICILS countries. Table 12.17 shows the item parameters for each of the two scales that were used to derive the IRT scale scores.

Figure 12.10: Confirmatory factor analysis of items measuring teachers' views on using ICT for teaching and learning at school

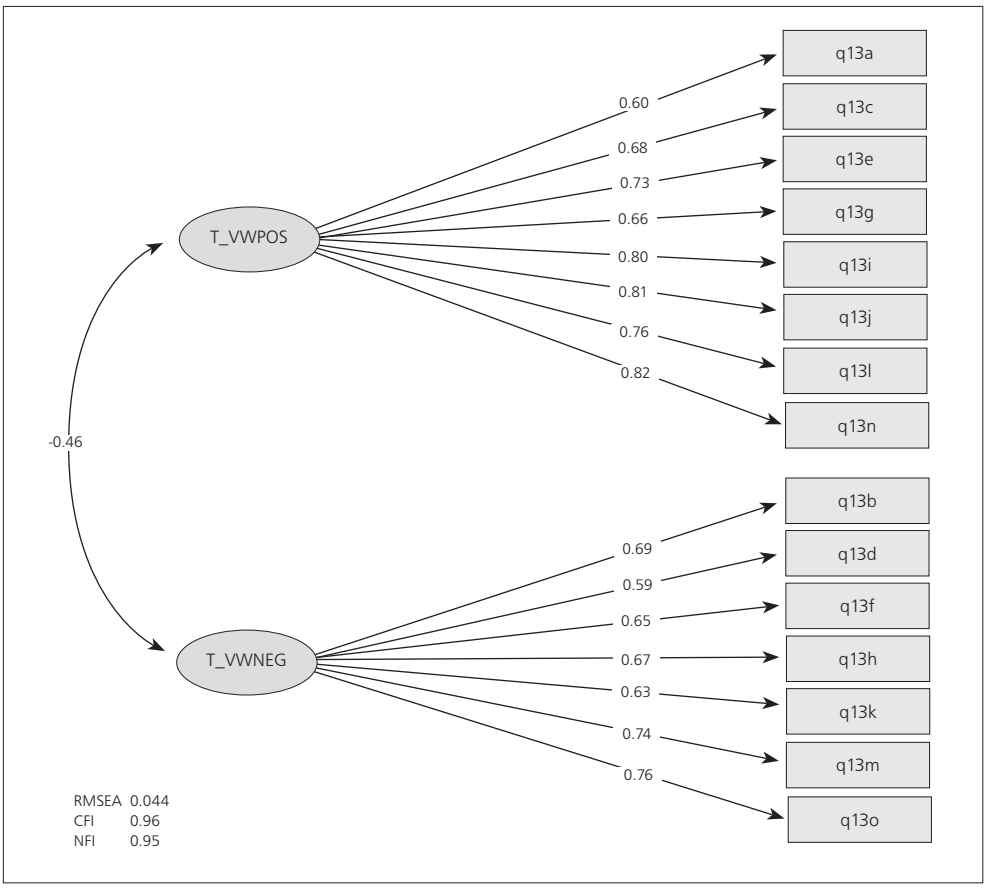


Table 12.16: Reliabilities for scales measuring teachers' views on using ICT for teaching and learning at school

Country	T_VWPOS	T_VWNEG
Australia	0.85	0.81
Chile	0.88	0.83
Croatia	0.85	0.81
Czech Republic	0.82	0.80
Denmark	0.82	0.77
Germany	0.82	0.78
Hong Kong SAR	0.80	0.76
Korea, Republic of	0.85	0.82
Lithuania	0.81	0.79
Netherlands	0.74	0.72
Newfoundland and Labrador, Canada	0.87	0.85
Norway (Grade 9)	0.81	0.78
Ontario, Canada	0.86	0.87
Poland	0.83	0.78
Russian Federation	0.84	0.80
Slovak Republic	0.82	0.78
Slovenia	0.81	0.76
Thailand	0.88	0.82
Turkey	0.86	0.82
Average reliability	**0.83**	**0.80**

Note: Benchmarking participants in italics.

Table 12.17: Item parameters for scales measuring teachers' views on using ICT for teaching and learning at school

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)
T_VWPOS	*To what extent do you agree or disagree with the following statements about using ICT in teaching and learning at school?*				
IT1G13A	Enables students to access better sources of information	-1.63	-2.50	-0.88	3.39
IT1G13C	Helps students to consolidate and process information more effectively	-0.94	-3.31	-0.57	3.88
IT1G13E	Helps students learn to collaborate with other students	0.22	-3.47	-0.45	3.91
IT1G13G	Enables students to communicate more effectively with others	0.66	-3.08	-0.17	3.26
IT1G13I	Helps students develop greater interest in learning	0.13	-3.16	-0.47	3.63
IT1G13J	Helps students work at a level appropriate to their learning needs	0.04	-3.65	-0.43	4.08
IT1G13L	Helps students develop skills in planning and self-regulation of their work	0.84	-3.82	-0.18	4.00
IT1G13N	Improves academic performance of students	0.68	-3.79	-0.19	3.98
T_VWNEG	*To what extent do you agree or disagree with the following statements about using ICT in teaching and learning at school?*				
IT1G13B	Results in poorer writing skills among students	-0.89	-2.81	0.21	2.59
IT1G13D	Only introduces organizational problems for schools	1.20	-3.14	0.96	2.18
IT1G13F	Impedes concept formation better done with real objects than computer images	0.12	-3.53	0.60	2.93
IT1G13H	Only encourages copying material from published internet sources	-0.42	-3.42	0.62	2.80
IT1G13K	Limits the amount of personal communication among students	-0.57	-3.04	0.35	2.69
IT1G13M	Results in poorer calculation and estimation skills among students	-0.23	-3.39	0.49	2.90
IT1G13O	Only distracts students from learning	0.79	-3.35	0.90	2.45

Teachers' views on the context for ICT use at their school

The teacher questionnaire included two questions asking teachers their views on the context for using ICT for teaching and learning at their school. We derived two scales from these questions. They were:

- Teachers' perspectives on the lack of computer resources at school (T_RESRC);
- Teachers' perspectives on collaboration between teachers in using ICT (T_COLICT).

Question 14 inquired about the extent to which teachers agreed or disagreed with statements about the use of ICT during teaching at their school. The four response options were "strongly agree," "agree," "disagree," and "strongly disagree." We used six of these eight items to derive a scale reflecting *teachers' perspectives on the lack of computer resources at school* (T_RESRC). The scale had an average reliability of 0.83, with coefficients ranging from 0.75 to 0.88 across the national samples. The higher values on the scale reflect greater deficiencies in resources available for using ICT as teaching tools.

Question 16 asked teachers to indicate the extent to which they agreed with five different practices and principles regarding the use of ICT for teaching and learning at their school. The response categories were "strongly agree," "agree," "disagree," and "strongly disagree." All items were used to derive a scale reflecting *teachers' perspectives on collaboration between teachers in using ICT* (T_COLICT). The scale had an average reliability of 0.79, and the coefficients ranged from 0.66 to 0.92. The higher scale values reflect greater collaboration between teachers.

Figure 12.11 depicts the results from a confirmatory factor analysis of the scaled items from the two questions. The two-factor model showed a satisfactory fit with a weak negative correlation between the two latent factors. Table 12.18 shows that the reliabilities for the two scales were satisfactory for most national samples, while Table 12.19 records the item parameters used for scaling these items with the Rasch partial credit model.

Figure 12.11: Confirmatory factor analysis of items measuring teachers' views on the context for ICT use at their school

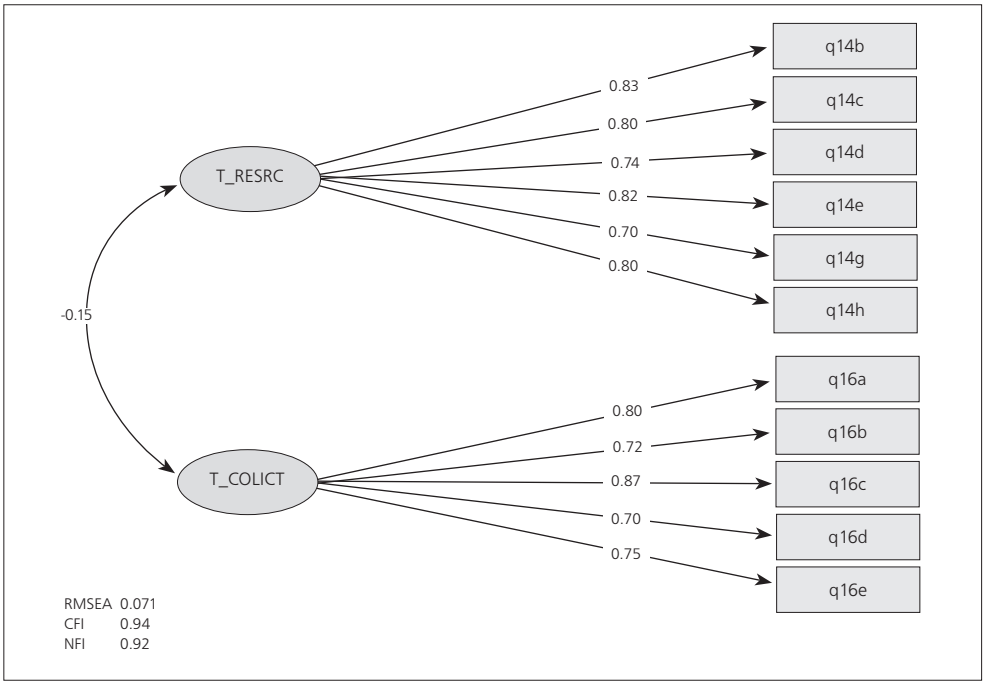


Table 12.18: Reliabilities for scales measuring teachers' views on the context for ICT use at their school

Country	T_RESRC	T_COLICT
Australia	0.85	0.81
Chile	0.85	0.83
Croatia	0.85	0.82
Czech Republic	0.79	0.75
Denmark	0.80	0.76
Germany	0.81	0.66
Hong Kong SAR	0.85	0.83
Korea, Republic of	0.84	0.80
Lithuania	0.80	0.75
Netherlands	0.75	0.75
Newfoundland and Labrador, Canada	0.84	0.77
Norway (Grade 9)	0.79	0.71
Ontario, Canada	0.85	0.82
Poland	0.84	0.82
Russian Federation	0.85	0.83
Slovak Republic	0.86	0.79
Slovenia	0.83	0.80
Thailand	0.88	0.92
Turkey	0.87	0.85
Average reliability	**0.83**	**0.79**

Note: Benchmarking participants in italics.

Table 12.19: Item parameters for scales measuring teachers' views on the context for ICT use at their school

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)
T_RESRC	*To what extent do you agree or disagree with the following statements about the use of ICT in teaching at your school?*				
IT1G14B	My school does not have sufficient ICT equipment (for example, computers)	-0.09	-2.12	0.26	1.86
IT1G14C	My school does not have access to digital learning resources	0.75	-2.81	0.72	2.08
IT1G14D	My school has limited connectivity (for example, slow or unstable speed) to the internet	-0.04	-2.36	0.34	2.02
IT1G14E	The computer equipment in our school is out-of-date	-0.03	-2.76	0.50	2.26
IT1G14G	There is not sufficient provision for me to develop expertise in ICT	-0.22	-2.92	0.42	2.50
IT1G14H	There is not sufficient technical support to maintain ICT resources	-0.37	-2.69	0.32	2.37
T_COLICT	*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?*				
IIT1G16A	I work together with other teachers on improving the use of ICT in classroom teaching	-0.45	-3.16	-0.34	3.50
IT1G16B	There is a common set of rules in the school about how ICT should be used in classrooms	0.12	-3.48	-0.05	3.53
IT1G16C	I systematically collaborate with colleagues to develop ICT-based lessons based on the curriculum	0.35	-3.55	0.06	3.49
IT1G16D	I observe how other teachers use ICT in teaching	-0.03	-2.94	-0.68	3.63
IT1G16E	There is a common set of expectations in the school about what students will learn about ICT	0.01	-3.39	-0.17	3.56

School questionnaires

ICT coordinators' reports on ICT resources at school

The ICT coordinator questionnaire asked a series of questions relating to the availability of different ICT resources (technology resources, software resources, and technology facilities) at the coordinators' respective schools. Items from these questions were used to derive a scale reflecting availability of ICT resources at school (C_ITRES).

In Questions 4, 5, and 6 of the ICT coordinator questionnaire, respondents were asked to indicate if the resources were "available" or "not available." Ten items from these three questions provided the basis for deriving the scale *availability of ICT resources at school* (C_ITRES). The higher scores on this scale indicate greater availability of resources.

Figure 12.12 presents the results of the confirmatory factor analysis. The one-dimensional model had a satisfactory fit. Table 12.20 shows the scale reliabilities (Cronbach's alpha), which had a relatively low average reliability (Cronbach's alpha) of 0.61. There were large variations in the coefficients across national samples. The range extended from 0.37 to 0.74. Table 12.21 shows the IRT item parameters that were used to derive the scale scores.

Figure 12.12: Confirmatory factor analysis of items measuring ICT coordinators' reports on ICT resources at school

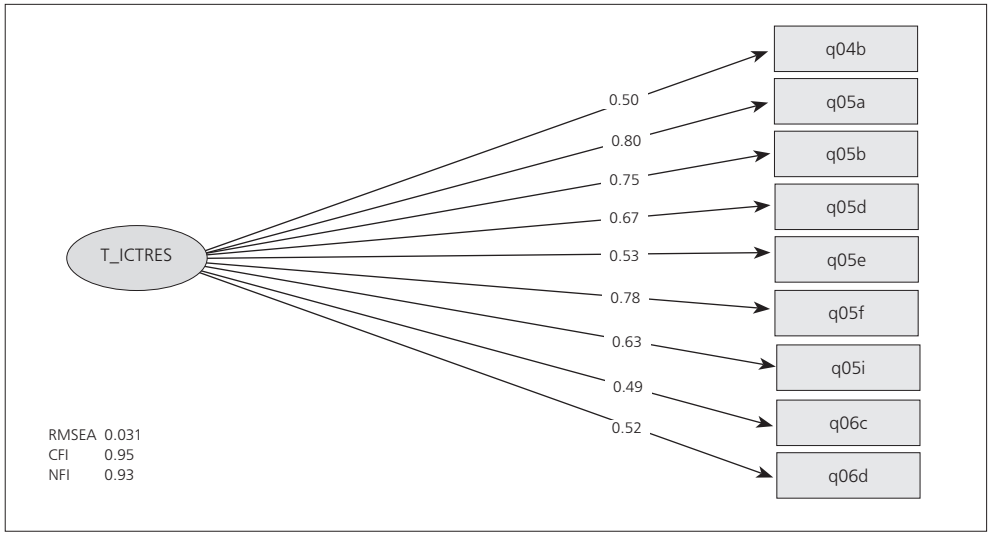


Table 12.20: Reliabilities for scale measuring ICT coordinators' reports on ICT resources at school

Country	C_ITRES
Australia	0.72
Buenos Aires, Argentina	0.67
Chile	0.59
Croatia	0.65
Czech Republic	0.49
Denmark	0.74
Germany	0.67
Hong Kong SAR	0.65
Korea, Republic of	0.70
Lithuania	0.52
Netherlands	0.37
Newfoundland and Labrador, Canada	0.54
Norway (Grade 9)	0.52
Ontario, Canada	0.73
Poland	0.62
Russian Federation	0.60
Slovak Republic	0.64
Slovenia	0.56
Switzerland	0.51
Thailand	0.65
Turkey	0.68
Average reliability	**0.61**

Note: Benchmarking participants in italics.

Table 12.21: Item parameters for scale measuring ICT coordinators' reports on ICT resources at school

Scale or Item	Question/Item Wording	Delta	Tau(1)
C_ITRES	*For each of the following technology resources please indicate their availability for teaching and/or learning*		
II1G04B	Interactive digital learning resources (for example, learning objects)	-1.02	
C_ITRES	*For each of the following software resources please indicate their availability for teaching and/or learning*		
II1G05A	Tutorial software or [practice programs]	-1.27	
II1G05B	Digital learning games	-0.52	
II1G05D	Multimedia production tools (for example, media capture and editing, web production)	-0.55	
II1G05E	Data logging and monitoring tools	1.03	
II1G05F	Simulations and modeling software	1.78	
II1G05I	Graphing or drawing software	-1.27	
C_ITRES	*For each of the following technology facilities please indicate their availability for teaching and/or learning at [target grade]*		
II1G06C	Space on a school network for students to store their work.	-0.02	
II1G06D	A school intranet with applications and workspaces for students to use (for example, [Moodle])	1.84	

ICT coordinators' perceptions of hindrances to ICT use at school

ICT coordinators were also asked (Question 13) to indicate to what extent ICT use in teaching and learning in their school was hindered by a range of different obstacles. Respondents were asked to indicate the extent of hindrance by selecting one of four response options ("a lot," "to some extent," "very little," and "not at all"). Responses to 10 of the 11 items provided the basis for deriving the following two scales:

- ICT use hindered in teaching and learning: lack of hardware (C_HINHW);
- ICT use hindered in teaching and learning: other obstacles (C_HINOTH).

The first scale (C_HINHW), with five items, had a reliability (Cronbach's alpha) of 0.80. The coefficients ranged from 0.64 to 0.88 across the participating countries. The second scale (C_HINOTH), also with five items, had an average reliability of 0.76 and national sample coefficients that spanned 0.35 to 0.88. The higher scores on these two scales represent greater extents of hindrance with regard to lack of hardware or other obstacles.

Figure 12.13 illustrates the results of the confirmatory factor analysis assuming a two-dimensional model with items from the two scales. The model fit was deemed satisfactory, and a moderate correlation was found between the two latent factors. Table 12.22 shows the (mostly satisfactory) scale reliabilities (Cronbach's alpha), and Table 12.23 records the IRT item parameters for each of the two scales.

Figure 12.13: Confirmatory factor analysis of items measuring ICT coordinators' perceptions of hindrances for ICT use at school

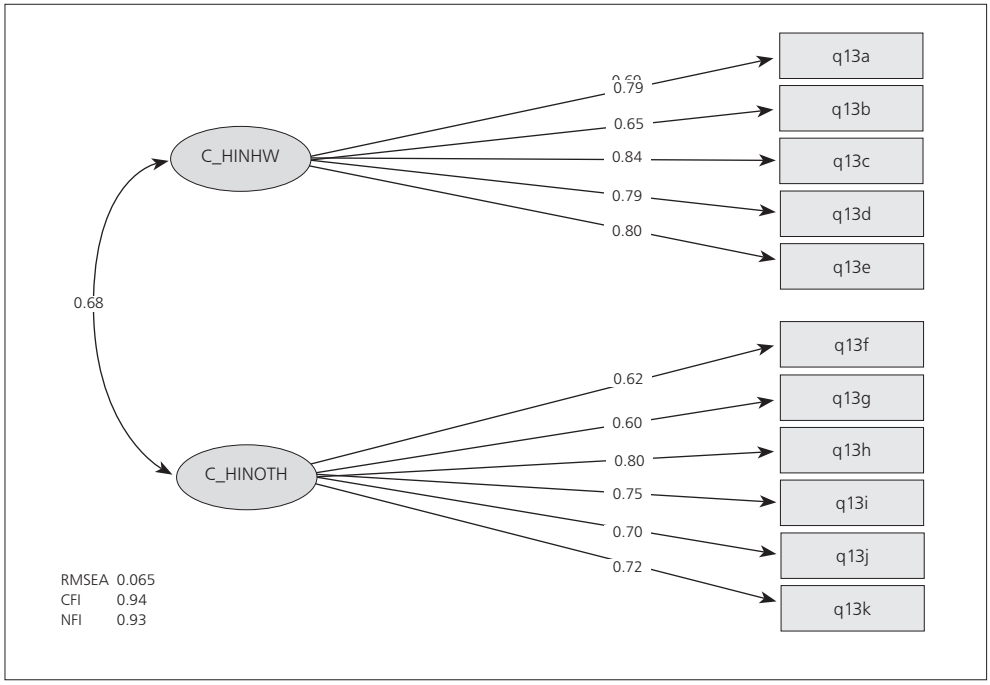


Table 12.22: Reliabilities for scale measuring ICT coordinators' perceptions of hindrances for ICT use at school

Country	C_HINHW	C_HINOTH
Australia	0.64	0.86
Buenos Aires, Argentina	0.82	0.75
Chile	0.84	0.80
Croatia	0.76	0.76
Czech Republic	0.78	0.70
Denmark	0.77	0.35
Germany	0.80	0.70
Hong Kong SAR	0.84	0.78
Korea, Republic of	0.88	0.88
Lithuania	0.73	0.75
Netherlands	0.80	0.85
Newfoundland and Labrador, Canada	0.78	0.86
Norway (Grade 9)	0.82	0.67
Ontario, Canada	0.81	0.85
Poland	0.84	0.78
Russian Federation	0.80	0.69
Slovak Republic	0.85	0.79
Slovenia	0.76	0.73
Switzerland	0.70	0.67
Thailand	0.82	0.85
Turkey	0.85	0.83
Average reliability	**0.80**	**0.76**

Note: Benchmarking participants in italics.

Table 12.23: Item parameters for scale measuring ICT coordinators' perceptions of hindrances for ICT use at school

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)	Tau(3)
C_HINHW	*To what extent is the use of ICT in teaching and learning in this school hindered by each of the following obstacles?*				
II1G13A	Too few computers connected to the internet	0.56	-0.76	-0.15	0.91
II1G13B	Insufficient internet bandwidth or speed	-0.02	-0.80	-0.18	0.98
II1G13C	Not enough computers for instruction	-0.23	-0.77	-0.29	1.06
II1G13D	Lack of sufficiently powerful computers	-0.28	-1.02	-0.23	1.25
II1G13E	Not enough computer software	-0.03	-1.63	-0.13	1.76
C_HINOTH	*To what extent is the use of ICT in teaching and learning in this school hindered by each of the following obstacles?*				
III1G13F	Lack of ICT skills among teachers	-0.04	-2.10	-0.47	2.57
II1G13G	Insufficient time for teachers to prepare lessons	-0.05	-1.91	-0.32	2.22
II1G13H	Lack of effective professional learning resources for teachers	-0.04	-2.17	-0.21	2.38
II1G13I	Lack of an effective online learning support platform	0.08	-1.72	-0.21	1.93
II1G13J	Lack of incentives for teachers to integrate ICT use in their teaching	0.05	-1.72	-0.34	2.07

School principals' perceptions of the importance of ICT at school

The questionnaire for schools included two questions measuring principals' perceptions of the importance assigned to ICT at their schools. These questions provided the basis for deriving the following two scales:

- Principals' perceptions of using ICT for educational outcomes (P_VWICT);
- Principals' perceptions of the ICT use expected of teachers: learning (P_EXPLRN).

Question 9 asked respondents to indicate the level of importance given to using ICT in their school in order to support different educational outcomes. The response categories were "very important," "somewhat important," and "not important." Four of the five items in the question were used to derive the first scale, *principals' perceptions of using ICT for educational outcomes* (P_VWICT), which had a reliability (Cronbach's alpha) of 0.75 across participating countries and national sample coefficients that ranged from 0.62 to 0.91. The higher scale values indicate higher levels of perceived importance.

Question 12 required respondents to indicate whether teachers in their school were expected to acquire knowledge and skills for 10 different activities. The response options were "expected and required," "expected but not required," and "not expected." We used seven of the 10 items to derive a scale reflecting *principals' perceptions of ICT use expected of teachers (learning)* (P_EXPLRN). The scale had an average reliability of 0.77; the coefficients ranged from 0.68 to 0.92 across the participating countries and benchmarking participants. The higher scale values indicate greater expectations of teachers.

Figure 12.14 illustrates the results of the confirmatory factor analysis. The two-dimensional model had a satisfactory model fit, and the results revealed a moderate correlation between the two latent factors. Table 12.24 shows the scale reliabilities (Cronbach's alpha) for the two scales. The reliabilities were satisfactory in most ICILS countries. Table 12.25 records the item parameters for each of the two scales used to derive the IRT scale scores.

Figure 12.14: Confirmatory factor analysis of items measuring principals' perceptions of the importance of ICT at school

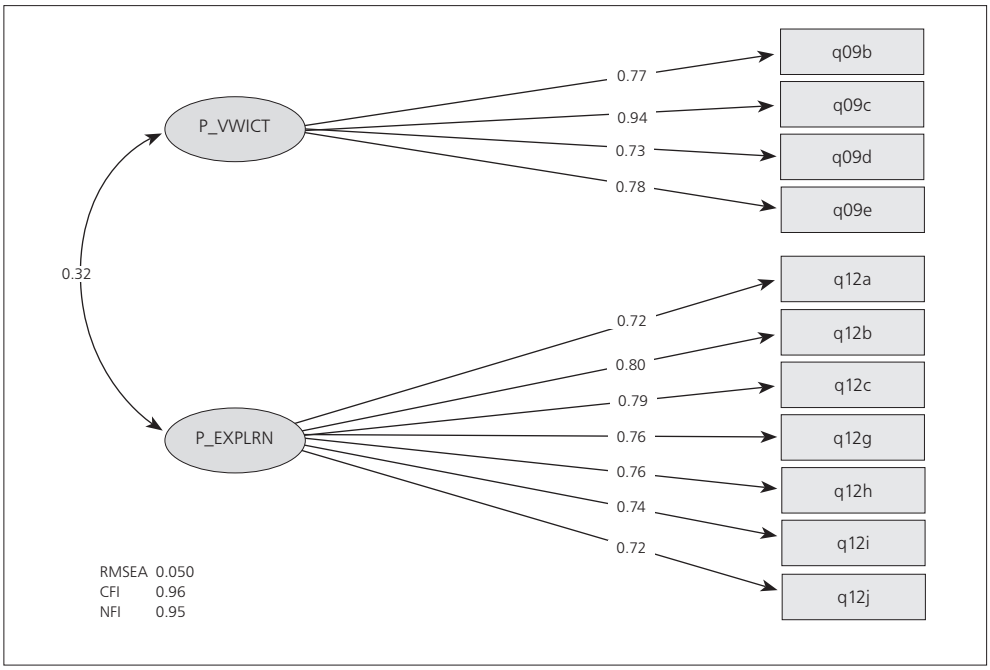


Table 12.24: Reliabilities for scale measuring school principals' perceptions of the importance of ICT at school

Country	P_VWICT	P_EXPLRN
Australia	0.70	0.69
Buenos Aires, Argentina	0.73	0.76
Chile	0.71	0.79
Croatia	0.80	0.76
Czech Republic	0.68	0.70
Denmark	0.66	0.80
Germany	0.75	0.76
Hong Kong SAR	0.68	0.78
Korea, Republic of	0.85	0.88
Lithuania	0.62	0.69
Netherlands	0.78	0.73
Newfoundland and Labrador, Canada	0.84	0.76
Norway (Grade 9)	0.78	0.78
Ontario, Canada	0.89	0.77
Poland	0.75	0.68
Russian Federation	0.71	0.73
Slovak Republic	0.65	0.72
Slovenia	0.74	0.79
Switzerland	0.70	0.82
Thailand	0.91	0.87
Turkey	0.84	0.9
Average reliability	**0.75**	**0.77**

Note: Benchmarking participants in italics.

Table 12.25: Item parameters for scale measuring school principals' perceptions of the importance of ICT at school

Scale or Item	Question/Item Wording	Delta	Tau(1)	Tau(2)
P_VWICT	*In your opinion, how important is the use of ICT in this school for each of the following outcomes of education?*			
IP1G09B	Using ICT for facilitating students' responsibility for their own learning	1.07	-2.29	2.29
IP1G09C	Using ICT to augment and improve students' learning	0.04	-2.63	2.63
IP1G09D	Developing students' understanding and skills relating to safe and appropriate use of ICT	-0.98	-2.16	2.16
IP1G09E	Developing students' proficiency in accessing and using information with ICT	-1.14	-2.28	2.28
P_EXPLRN	*Are teachers in your school expected to acquire knowledge and skills in each of the following activities?*			
IP1G12A	Integrating web-based learning in their instructional practice	-0.66	-2.52	2.52
IP1G12B	Using ICT-based forms of student assessment	0.43	-1.88	1.88
IP1G12C	Using ICT for monitoring student progress	0.11	-1.71	1.71
IP1G12G	Integrating ICT into teaching and learning	-2.12	-2.44	2.44
IP1G12H	Using subject-specific learning software (for example, tutorials, simulation)	-0.29	-2.17	2.17
IP1G12I	Using e-portfolios for assessment	1.85	-1.55	1.55
IP1G12J	Using ICT to develop authentic (real-life) assignments for students	0.70	-2.30	2.30

School principals' views of ICT priorities at school

Question 16 of the ICILS principal questionnaire asked respondents to indicate the priority the school gave to ways of facilitating ICT use in teaching and learning. The response categories were "high priority," "medium priority," "low priority," and "not a priority." The responses to this question provided data for deriving the following two scales:

- Principals' views of priorities for facilitating use of ICT: hardware (P_PRIORH);
- Principals' views of priorities for facilitating use of ICT: support (P_PRIORS).

Exploratory factor analyses indicated a two-dimensional structure for the items pertaining to this question. The first scale, reflecting *principals' perceived priorities regarding hardware* (P_PRIORH), was derived from the first three question items and had an average reliability (Cronbach's alpha) of 0.76, with national coefficients ranging from 0.46 to 0.89. The remaining seven items provided the basis for deriving the second scale, which reflected *principals' perceived priorities for improving support for ICT use* (P_PRIORS). This scale had an average reliability of 0.82 and coefficients ranging from 0.72 to 0.94. The higher scores on these two scales represent higher levels of priority for each aspect, as perceived by the school principals.

Figure 12.15 provides the results of the confirmatory factor analysis. Here we can see that the two-dimensional model had a satisfactory fit. The results also suggested a relatively high positive correlation between the two latent factors (0.60). Table 12.26 shows the scale reliabilities (Cronbach's alpha) for the two scales, which were satisfactory in most ICILS countries. Table 12.27 shows the IRT item parameters that we used to derive the final scale scores for each of the two scales.

Figure 12.15: Confirmatory factor analysis of items measuring principals' views of ICT priorities at school

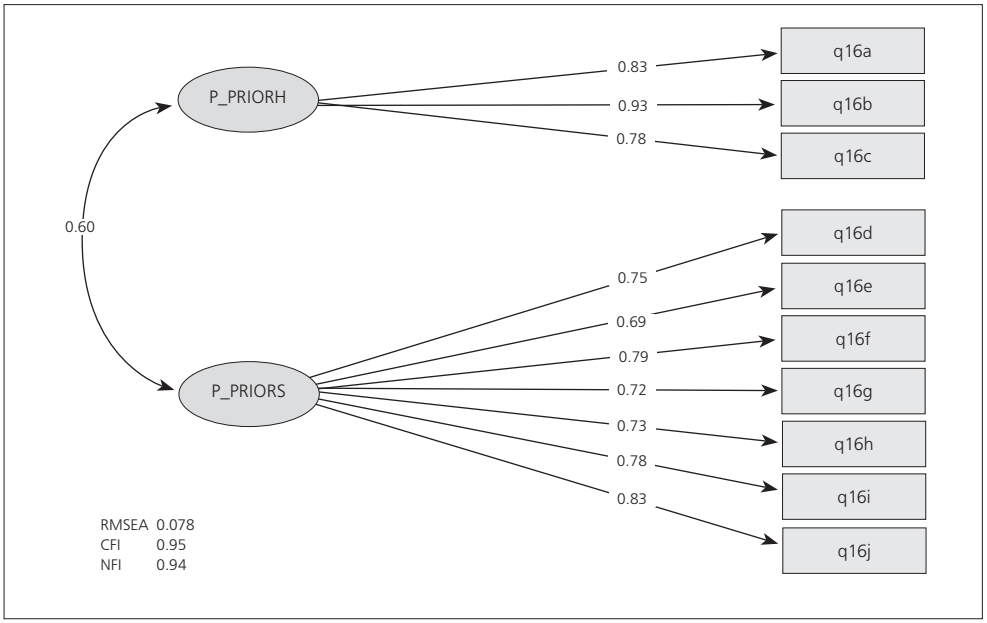


Table 12.26: Reliabilities for scale measuring school principals' views of ICT priorities at school

Country	P_PRIORH	P_PRIORS
Australia	0.78	0.82
Buenos Aires, Argentina	0.83	0.89
Chile	0.85	0.89
Croatia	0.75	0.80
Czech Republic	0.84	0.75
Denmark	0.70	0.72
Germany	0.81	0.81
Hong Kong SAR	0.76	0.79
Korea, Republic of	0.89	0.90
Lithuania	0.81	0.81
Netherlands	0.73	0.79
Newfoundland and Labrador, Canada	0.81	0.87
Norway (Grade 9)	0.84	0.74
Ontario, Canada	0.72	0.85
Poland	0.46	0.88
Russian Federation	0.75	0.83
Slovak Republic	0.66	0.81
Slovenia	0.72	0.72
Switzerland	0.73	0.73
Thailand	0.78	0.94
Turkey	0.73	0.86
Average reliability	**0.76**	**0.82**

Note: Benchmarking participants in italics.

Table 12.27: Item parameters for scale measuring school principals' views of ICT priorities at school

Scale or Item	Question/Item Wording	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?*				
IP1G16A	Increasing the numbers of computers per student in the school	0.14	-1.53	-0.52	2.06
IP1G16B	Increasing the number of computers connected to the internet	-0.18	-1.34	-0.18	1.53
IP1G16C	Increasing the bandwidth of internet access for the computers connected to the internet	0.05	-1.56	-0.58	2.14
P_PRIORS	*At your school, what priority is given to the following ways of facilitating the use of ICT in teaching and learning?*				
IP1G16D	Increasing the range of digital learning resources	-0.58	-1.74	-0.68	2.42
IP1G16E	Establishing or enhancing an online learning support platform	0.49	-1.85	-0.17	2.02
IP1G16F	Providing for participation in professional development on pedagogical use of ICT	-0.59	-1.94	-0.40	2.35
IP1G16G	Increasing the availability of qualified technical personnel to support the use of ICT	0.14	-1.50	-0.38	1.88
IP1G16H	Providing teachers with incentives to integrate ICT use in their teaching	0.15	-1.10	-0.44	1.54
IP1G16I	Providing more time for teachers to prepare lessons in which ICT is used	0.70	-1.37	-0.53	1.90
IP1G16J	Increasing the professional learning resources for teachers in the use of ICT	-0.32	-1.59	-0.67	2.26

Summary

ICILS derived two different types of indices from the questionnaires administered to students, teachers, and schools. Several simple indices were constructed by either recoding values, combining separate variables, or using arithmetic. A second type of index was derived through scaling procedures.

Item response modeling (applying the Rasch partial credit model) provided an adequate tool for deriving 10 international student questionnaire scales, nine teacher questionnaire scales, and seven school questionnaire scales. A composite index reflecting socioeconomic background was derived using principal component analysis of three home background indicators, namely, parental occupation, parental education, and home literacy resources.

Generally, the scales used in ICILS had sound psychometric properties, such as high reliabilities. Confirmatory factor analyses showed satisfactory model fit for the measurement models underpinning the scaling of the questionnaire data.

References

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. (1993). (Eds.). *Testing structural equation models*. Newbury Park, CA: Sage Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Pscyhometrika*, *16*(3), 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., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA International Computer and Information Literacy Study international report*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491.

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.

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: Author.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extension*s. Los Angeles, CA: Sage Publications.

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), 130–149.

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, NY: 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 online at 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*. 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., & 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).

Tucker, L., & MacCallum, R. (1997). *Exploratory factor analysis*. Unpublished manuscript. Retrieved from http://www.unc.edu/~rcm/book/factornew.htm

UNESCO. (2006). *International Standard Classification of Education: ISCED 1997*. Montreal, Quebec, Canada: UNESCO-UIS.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427−450.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest: General item response modelling software*. Camberwell, Victoria, Australia: Australian Council for Educational Research (ACER).

CHAPTER 13:

The Reporting of ICILS Results

Wolfram Schulz

Introduction

This chapter describes the procedures that were used to report the population estimates in the ICILS 2013 publications. It starts with a description of the replication methodology used to estimate sampling variance, followed by an outline of how the imputation variance of students' computer and information literacy (CIL) was computed. The subsequent section outlines the procedures for conducting significance tests of differences between country and subsample means or percentages.

The international report on ICILS data (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014) presented results from both single-level multiple regression models and hierarchical linear modeling to explain students' CIL scores. Predictor variables were student background variables, student characteristics, and students' familiarity with ICT, as well as school context factors. The chapter includes technical descriptions of the multiple regression analyses and hierarchical models explaining CIL that were included in the ICILS international report. The final section of the chapter outlines how missing data were treated during the multivariate analyses of the ICILS 2013 data.

Estimation of sampling variance

ICILS employed two-stage cluster sampling procedures to obtain the student as well as the teacher samples. During the first stage, schools were sampled from a sampling frame with a selection probability proportional to their size. During the second stage, students enrolled in the target grade (typically Grade 8) were randomly sampled from within schools. Cluster sampling techniques permit an efficient and economic data collection. However, because these ICILS samples were not simple random samples, the usual formulae used to obtain standard errors for population estimates were not appropriate.

Replication techniques provide tools with which to estimate the correct sampling variance on population estimates (Gonzalez & Foy, 2000; Wolter, 1985). The ICILS research team applied the jackknife repeated replication technique (JRR) to compute standard errors of population means, percentages, regression coefficients, and any other population statistic.

In a nutshell, "replication" methodology involves selecting subsamples from the achieved sample and comparing the variability of these replicate samples with the original sample, which makes it possible to obtain estimates of sampling variance. In JRR, the selection of subsamples from the sample is technically done by altering the original sampling weights.

As a first step, the JRR method for stratified samples requires pairing primary sampling units (PSUs)—in ICILS, schools—with pseudo-strata.[1] Because assignment of schools to these "jackknife zones" needed to be consistent with the sampling frame from which

1 In the Russian Federation, regions were primary sampling units. Within those regions selected with certainty, schools were paired as in all other countries. Otherwise, regions were paired.

they were sampled, we constructed jackknife zones within explicit strata. In order to account for the effects of implicit stratification on sampling error, we paired schools adjacent to each other in the sorted sample into one jackknife zone. When faced with occurrences of an odd number of schools within an explicit stratum, we randomly divided students from the remaining school into two halves, thereby forming an extra jackknife zone of two "quasi-schools."

Because the minimum school sample size in ICILS was 150 schools, we set the maximum number of pseudo-strata to 75, which meant that each of the countries participating in ICILS had up to 75 jackknife zones. In countries with a large number of participating schools, we combined some schools into bigger "pseudo-schools" in order to keep the total number of zones to 75. In some countries, fewer zones were needed. Table 13.1 shows the range of jackknife zones for the student, school, and teacher samples used in each participating country.

The next step involved assigning a multiplication value to each school. Within each of the jackknife zones, we randomly assigned one school a value of 2 and the other school a value of 0. Based on these values, we created 75 so-called "replicate weights" by multiplying the final weight of a case (in ICILS, students, teachers, or schools) with the multiplication value following a specific pattern.

This multiplication was performed for the first replicate weight only for cases belonging to jackknife zone 1; all other cases kept their original final weight. Hence, in replicate weight 1 in jackknife zone 1, some participants received twice their original weight while others had their replicate weight value set to 0. In the second replicate weight, the

Table 13.1: Number of jackknife zones in national samples

Country	Student Data	Teacher Data	School Data
Australia	75	75	75
Chile	75	75	75
Croatia	75	75	75
Czech Republic	75	75	75
Denmark	52	41	46
Germany	70	62	67
Hong Kong SAR	59	54	58
Korea	75	75	75
Lithuania	75	75	75
Netherlands	61	49	52
Norway (Grade 9)	71	60	65
Poland	75	75	75
Russian Federation	62	62	62
Slovak Republic	75	75	75
Slovenia	75	75	75
Switzerland	50	N/A	41
Thailand	75	75	75
Turkey	71	75	75
Benchmarking participants			
City of Buenos Aires, Argentina	35	N/A	35
Newfoundland and Labrador, Canada	75	75	53
Ontario, Canada	75	64	75

multiplication was performed only for participants in jackknife zone 2, and so on. This process meant that estimating any population characteristic using one of these replicate weights would lead to some cases of the original sample contributing twice and others not contributing at all.

Table 13.2 illustrates this procedure through a simple example featuring 24 students from six different schools (A–F) paired into three sampling zones.

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

For each country sample, we computed 75 replicate weights regardless of the number of jackknife zones. In countries with fewer jackknife zones, the remaining replicate weights were equal to the original final weights (student, teacher, or school) and therefore did not contribute to the sampling variance estimate.

Estimating the sampling variance of a statistic μ involves computing it once with the final weights for the original sample and then with each of the 75 replication weights separately. The sampling variance $SV\mu$ estimate is computed using the formula

$$SV_{\mu} = \sum_{i=1}^{75} [\mu_i - \mu_s]^2$$

where 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 $SE\mu$ for statistic μ is computed as:

$$SE_{\mu} = \sqrt{SV_{\mu}}$$

The computation of sampling variance using jackknife replication can be obtained for any statistic, including means, percentages, standard deviations, correlations, regression coefficients, and mean differences. Standard statistical software generally does not include procedures for replication techniques.[2]

For the jackknife replication of ICILS data, we applied tailored SPSS software macros. Analysts can replicate these results by using the IEA IDB Analyzer, which is generally recommended as a tool for analyzing IEA data.[3] Alternatively, analysts can use other specialized software, such as WESVAR (Westat, 2007), or tailored applications, such as the SPSS Replicates Module developed by the Australian Council for Educational Research (ACER).[4]

Estimation of imputation variance for CIL scores

When estimating standard errors for test scores reflecting CIL, it is important to take the imputation variance into account because this indicates the level of precision in the measurement of the latent trait (see Chapter 11 for a description of the scaling methodology for ICILS test items). Population statistics for CIL scores should therefore always be estimated via use of all five plausible values.

If θ is the student's international CIL score and μ_θ^p is the statistic of interest computed based on each of the P plausible values, 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 makes it possible to estimate the amount of error associated with the measurement of CIL. The measurement variance or imputation variance IV_P is computed as

$$IV_P = \frac{1}{P-1} \sum_{p=1}^{P} (\mu_\theta^p - \mu_\theta)^2$$

The estimate of the total variance TV_μ, consisting of sampling variance and imputation variance, can be computed as

$$TV_\mu = SV_\mu + \left(1 + \frac{1}{P}\right) IV_\mu$$

The estimate of the final standard error SE_μ is equal to

$$SE_\mu = \sqrt{TV_\mu}$$

2 Procedures for replication techniques are now built into newer versions of the statistical software packages SAS and MPlus.

3 The IDB Analyzer is a plug-in for the Statistical Package for the Social Sciences (SPSS). It allows users to combine and analyze data from all IEA large-scale studies. The application can be downloaded at http://www.iea.nl/iea_studies_datasets.html

4 The module is an add-in component running under SPSS. It offers some features for applying different replication methods when estimating sampling and imputation variance. The application can be downloaded at https://mypisa.acer.

Table 13.3 shows the average scale scores as well as their sampling and overall standard errors. It also records the number of students assessed in each country. The comparison between sampling error and combined standard error shows that most of the error was due to sampling and that only a small proportion of it could be attributed to measurement error.

Table 13.3: National averages for CIL with sampling error, combined standard error, and the number of assessed students

Country	Average CIL Score	Sampling Error	Combined Standard Error	Number of Assessed Students
Australia	542	2.24	2.28	5326
Chile	487	3.11	3.14	3180
Croatia	512	2.66	2.86	2850
Czech Republic	553	1.99	2.05	3066
Denmark*	542	3.24	3.49	1767
Germany	523	2.31	2.41	2225
Hong Kong SAR*	509	7.42	7.44	2089
Korea, Republic of	536	2.61	2.67	2888
Lithuania	494	3.54	3.60	2756
Netherlands*	535	4.63	4.68	2197
Norway (Grade 9)	537	2.32	2.39	2436
Poland	537	2.36	2.42	2870
Russian Federation	516	2.68	2.82	3626
Slovak Republic	517	4.47	4.55	2994
Slovenia	511	2.15	2.18	3740
Switzerland*	526	4.51	4.58	3225
Thailand	373	4.63	4.68	3646
Turkey	361	4.93	4.97	2540
Benchmarking participants				
Buenos Aires, Argentina*	450	8.57	8.59	1076
Newfoundland and Labrador, Canada	528	2.57	2.81	1556
Ontario, Canada	547	3.16	3.22	3377

Note: * Countries not meeting sample participation requirements.

Reporting of differences

The international report includes significance tests for:

- Differences in population estimates across countries (multiple comparisons);
- Differences between a country average and the international average;
- Differences in population estimates among subgroups within countries.

We considered differences between two score averages (or percentages) a and b as significant ($p < 0.05$) when the absolute value of the test statistic t was greater than the critical value, 1.96. The test statistic t is calculated 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 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, as is the case with countries or explicit strata within countries. 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 replication that involved the following formula:

$$SE_{dif\_ab} = \sqrt{\sum_{i=1}^{75} ((a^i - b^i) - (a - b))^2}$$

Here, a and b represent the weighted averages (or percentages) in each of the two subgroups for the fully weighted sample, while a^i and b^i are the weighted averages for the replicate samples.

In the case of differences in CIL scores between dependent subsamples, we calculated the standard error of the differences with a specified number ($P = 5$) of plausible values by using this formula:

$$SE_{dif\_ab} = \sqrt{\left[\frac{\sum_{p=1}^{P}\left(\sum_{i=1}^{75}((a_p^i - b_p^i) - (a_p - b_p))^2\right)}{P}\right] + \left[\left(1 + \frac{1}{P}\right)\frac{\sum_{p=1}^{P}((a_p - b_p) - (\bar{a}_p - \bar{b}_p))^2}{P-1}\right]}$$

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.

When comparing the country means c with the overall ICILS average i, we needed 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 average and the 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 the k^{th} of the N participating countries. This formula for determining the statistical significance of differences due to sampling error between countries and the ICILS averages of the questionnaire scales was used throughout the ICILS international report (Fraillon et al., 2014).

While the above formula was sufficient for the questionnaire scale scores, we found it necessary to also take the imputation component of standard errors for countries into

account when comparing the test score averages of a country with the overall ICILS average. The imputation variance component of standard errors $SE^2_{i\,dif\_ic}$ was given as

$$SE^2_{i\_dif\_ic} = \sqrt{\left(1+\frac{1}{P}\right)\ var(d_1,\ldots,d_p,\ldots,d_5)}$$

where d_p is the difference between the overall ICILS mean and the country mean for the plausible value *p*.

The final standard error ($SE_{a\_dif\_ic}$) of the difference between ICILS country test scores and the ICILS average was computed as

$$SE_{a\_dif\_ic} = \sqrt{SE^2_{dif\_ic} + SE^2_{i\_dif\_ic}}$$

Multiple regression modeling

The international report on ICILS presented multiple regression models in order to explain the variation not only in students' CIL scores associated with their personal and home background, but also in teachers' emphasis on ICT for teaching with their confidence in using ICT, views of ICT use, collaboration with other teachers at school, and lack of ICT resources at school.

When conducting multiple regression models, analysts regress a criterion variable Y_i on a set of *k* predictors $X_{li} \ldots X_{ki}$, with α being the intercept, ε_i the unexplained part of the model (residual), and β_k the regression coefficients:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots + \beta_k X_{ki} + \varepsilon_i$$

We estimated multiple regression models in order to review the influence of family background. Unstandardized regression coefficients and the variance explained by the model were reported in the international report so as to show the effect of each predictor and the overall explanatory power of the model. Jackknife replication applied via tailored SPSS macros allowed us to estimate appropriate standard errors for the multiple regression model parameters.

When analyzing the influence of students' personal and home factors on their CIL, we wanted to review the relative importance of the different sets of predictor variables. In order to estimate the unique contribution of each set of predictors to the variance explanation of the model as well as the proportion of variance explained by more than one set of predictors, we estimated a series of *j* different linear regression models. For each set of *j* with *m* predictor variables, we omitted one predictor set from the model. The difference in variance explanation for the full model and the model without a certain set of predictors showed the unique contribution this set of factors made with respect to explaining variance in the criterion variable. The variance uniquely explained for predictor variable set *j* ($r^2_{u\_j}$) was obtained as

$$r^2_{u\_j} = r^2_n - r^2_{n-m}$$

where r^2_n is the *R* square for the full model and r^2_{n-m} is the *R* square for the regression model without the *m* variables in predictor block *j*.

The unique contribution of predictor set *j* to the explained variance in the predictor variable Y_i can be expressed in percentages as

$$UVC_j = r^2_{u\_j} * 100$$

The joint explained variance contribution reflects the proportion of variance explained by more than one k set of predictors. The proportion of variance explained by more than one set of predictors JVC_j was computed as

$$JVC_j = (r_n^2 * 100) - \sum_{k=1}^{k} UVC_j$$

Hierarchical linear modeling

When analyzing the influence of student background, ICT familiarity, and school factors on students' CIL, we knew it was important to take school or classroom context effects into account. In order to achieve this, we used hierarchical (or multilevel) linear regression models (Raudenbush & Bryk, 2002) in which students were nested within schools.

A hierarchical regression model with i students nested in j clusters (schools) can be estimated as

$$Y_{ij} = \alpha_j + X_{ij}^n \beta_{ij} + X_j^m \beta_j + U_{0j} + \varepsilon_{ij}$$

where Y_{ij} is the criterion variable, X_{ij}^n is a vector of student-level variables with its corresponding vector of regression coefficients β_{ij}, and X_j^m 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 (here, 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 is 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. Thus:

$$Y_{ij} = \alpha_j + U_{0j}^{null} + \varepsilon_{ij}^{null}$$

The residual term U_{0j}^{null} provides an estimate of the variance in Y_{ij} between j clusters, and ε_{ij}^{null} is an estimate of the variance between i students within clusters. The intraclass correlation IC, which reflects the proportion of variance between clusters (in our case, schools), can be computed from these estimates as

$$IC = \frac{U_{0j}^{null}}{U_{0j}^{null} + \varepsilon_{ij}^{null}}$$

The "null" model—that is, the model obtained from including only students without missing data after completing "missing treatment" (see the section on missing treatment below)—provided a baseline model for the ICILS multilevel analysis. Comparisons between the variance estimates at each level of the baseline model U_{0j}^{null} and those estimates with predictor variables (U_{0j}) enabled us to compute explained variance at the school level EV_j as

$$EV_j = (1 - \frac{U_{0j}}{U_{0j}^{null}}) * 100$$

while the explained variance at the student level EV_{ij} was equal to

$$EV_{ij} = (1 - \frac{\varepsilon_{ij}}{\varepsilon_{ij}^{null}}) * 100$$

We did not use the JRR method to estimate the standard errors of the multilevel model parameters because this analysis technique had already taken the hierarchical structure of the cluster sample into account. However, reported standard errors accounted for imputation error. We accordingly weighted data (with normalized school-level and [within-school] student-level weights) following a recommended procedure for the analysis of IEA data (see Rutkowski, Gonzalez, Joncas, & von Davier, 2010).

We conducted the analysis using the hierarchical modeling tool in the software package MPlus 7 (Muthén & Muthén, 2012). The results of our hierarchical data analysis using all three data sources (student, school, and teacher data) from the education systems that did not meet IEA sample participation requirements (i.e., Hong Kong SAR and Denmark)were reported in separate sections of the tables in the international report. We excluded three education systems with extremely low teacher participation rates from these analyses. These systems were the city of Buenos Aires in Argentina, the Netherlands, and Switzerland.

We included teacher survey data in the analysis by computing (weighted) average scale scores at the school level. We then used the school averages of the scale scores reflecting teachers' views on resource limitations restricting use of ICT for teaching and learning as the school-level predictor variable.

Missing data treatment in ICILS multivariate analyses

When conducting multivariate analysis, researchers usually find that missing data problems tend to be more salient in this type of analysis than in other forms with only one or two variables. A larger number of cases might have to be excluded if the analysis uses only those records that have complete information for all variables. Generally, in ICILS, there were two possible sources of missing data: (i) completely missing questionnaire data either for the student or their school, and (ii) missing data for individual variables.

The proportion of missing data at the student level was relatively low and allowed us to exclude students with missing data from the multiple regression analyses without losing much information (as would be the case, for example, when 93% of students have valid data for all student background predictor variables during prediction of CIL). However, multilevel analysis of CIL included school-level variables with higher percentages of missing data and therefore required a different treatment method.

We excluded from the multilevel analyses the small proportion of students without student questionnaire data, as well as those who had missing values on one or more of the student-level variables. However, in order to account for higher proportions of missing responses from the ICT coordinator survey and the teacher survey in some participating countries, we included a "dummy variable adjustment" for these data in the analysis (see Cohen & Cohen, 1975). We then assigned mean or median values to schools with missing data and added dummy indicator variables (with 1 indicating a missing value and 0 nonmissing values) to the analysis.

Given that the information from ICT coordinators tended to be either not missing or missing (in almost all cases) for both of the two predictor variables from this survey (the variables were *availability of ICT resources for teaching and learning* and *school experience with using ICT for teaching and learning*), we created only one missing indicator to indicate missing ICT coordinator data for both variables. We used another

missing indicator for missing teacher survey data on the predictor variable ICT resource limitations for teaching and learning.

Table 13.4 shows the unstandardized regression coefficients for the two missing indicators for the two hierarchical linear models reported in the international report (Fraillon et al., 2014, p. 225ff). We found no significant effects of missing indicators on CIL in most countries. The Republic of Korea and the Slovak Republic recorded positive coefficients between missing ICT coordinator and teacher data on CIL. Thus, in these countries, schools with missing school and teacher data tended to have the higher average CIL scores. Table 13.4 also shows that in Australia, the Model 2 missing ICT coordinator data showed a negative association. A negative coefficient was also recorded for missing teacher survey information in Norway. Generally, though, we found no consistent associations between the indicators of missing school-level data and CIL.

Table 13.4: Coefficients of missing indicators in multilevel analysis of ICILS data

Country	Indicator Variables of Missing...			
	School data		Teacher data	
	Model 1	*Model 2*	*Model 1*	*Model 2*
Australia	-12.8 (10.5)	**-16.8** (8.2)	12.2 (11.8)	2.8 (8.9)
Chile	25.3 (19.9)	8.8 (9.8)	25.3 (19.9)	8.8 (9.8)
Croatia	-8.5 (13.4)	-10.9 (9.5)	-8.5 (13.4)	-10.9 (9.5)
Czech Republic	0.2 (1.4)	0.3 (1.4)	0.2 (1.4)	0.3 (1.4)
Denmark	-4.1 (10.3)	-12.0 (8.7)	-2.7 (9.1)	-6.1 (6.7)
Germany	-8.3 (19.5)	-14.2 (11.7)	2.9 (14.6)	1.9 (8.2)
Hong Kong SAR	9.2 (16.3)	9.8 (16.0)	-33.2 (17.6)	-29.9 (17.9)
Korea	**7.3** (2.7)	**5.4** (2.1)	**7.3** (2.7)	**5.4** (2.1)
Lithuania	-15.5 (13.8)	-13.9 (14.3)	-15.5 (13.8)	-13.9 (14.3)
Norway (Grade 9)	7.6 (7.0)	4.5 (5.5)	**-18.3** (7.8)	**-15.0** (5.6)
Poland	7.9 (6.5)	-4.5 (5.8)	7.9 (6.5)	-4.5 (5.8)
Russian Federation	18.9 (13.6)	15.7 (12.3)	4.7 (16.0)	3.2 (15.3)
Slovak Republic	**71.8** (22.4)	**45.1** (21.9)	**71.8** (22.4)	**45.1** (21.9)
Slovenia	6.4 (10.3)	7.7 (8.3)	13.7 (9.5)	4.5 (6.6)
Thailand	8.7 (22.2)	10.9 (18.2)	-44.2 (30.3)	-40.9 (26.5)
Turkey	-26.6 (28.6)	-28.4 (26.5)	-26.6 (28.6)	-28.4 (26.5)
Benchmarking participants				
Newfoundland and Labrador, Canada	-7.1 (10.3)	-7.7 (9.8)	-24.4 (13.2)	-14.2 (12.2)
Ontario, Canada	-7.2 (6.6)	-9.2 (6.6)	-8.5 (7.2)	-3.5 (7.3)

Note: Statistically significant ($p < 0.05$) coefficients in bold.

Table 13.5 shows (for the countries included in the multilevel analysis) the numbers of students assessed in ICILS and the respective percentages of students included in the analysis after we had completed the missing treatment. Overall, 98 percent of assessed students were included in the analysis after treatment. Inclusion percentages across the participating countries ranged from 93 percent in Germany to (almost) 100 percent in the Republic of Korea.

Table 13.5: ICILS students included in multilevel analysis of CIL

Country	Total Number of Assessed Students	Total Number of Students in Analysis	Percentage of Students in Analysis
Australia	5326	5237	98.3
Chile	3180	3102	97.5
Croatia	2850	2827	99.2
Czech Republic	3066	3030	98.8
Denmark	1767	1717	97.2
Germany	2225	2068	92.9
Hong Kong SAR	2089	2041	97.7
Korea, Republic of	2888	2874	99.5
Lithuania	2756	2670	96.9
Norway	2436	2397	98.4
Poland	2870	2829	98.6
Russian Federation	3626	3559	98.2
Slovak Republic	2994	2934	98.0
Slovenia	3740	3671	98.2
Thailand	3646	3525	96.7
Turkey	2540	2400	94.5
Benchmarking participants			
Newfoundland and Labrador, Canada	1556	1497	96.2
Ontario, Canada	3377	3255	96.4
Overall sample	**52,932**	**51,633**	**97.5**

Summary

In order to report sampling errors in ICILS reports, the jackknife repeated replication technique (JRR) was used. Employing plausible value methodology to derive CIL scores allowed the estimation of imputation variance in addition to the sampling variance.

Different types of significance test were applied during comparison of means or percentages between participating countries, with the ICILS average, or between subgroups within the sample.

The ICILS international report included multiple regression models as well as multilevel models. When applying (two-level) hierarchical linear modeling, the ICILS research team compared two different models—with and without controls for student characteristics and socioeconomic background. Estimates of explained variance were computed separately at the student and school levels.

While students with missing data could be excluded from the multiple regression analysis without losing too much information, higher proportions of missing data from some school-level data had to be taken into account during the multilevel analyses of variation in CIL. We addressed this problem by adding missing indicators and substituting missing values with modes or means.

Secondary analysts should conduct multivariate analysis of ICILS data by taking potential missing data problems into account and exploring possibilities for applying more advanced methods, including imputation procedures.

References

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

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* (pp. 203–222). Chestnut Hill, MA: Boston College.

Muthén, L. K., & Muthén, B. O. (2012). *MPlus: Statistical analysis with latent variables. User's guide* (Version 7). Los Angeles, CA: Muthén & Muthén.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publishers.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.

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.

Appendices

APPENDIX A:

Organizations and individuals involved in ICILS

International study center

The international study center is located at the Australian Council for Educational Research (ACER). Center staff at ACER were responsible for designing and implementing the study in close cooperation with the IEA Data Processing and Research Center (DPC) in Hamburg, Germany, and the IEA Secretariat in Amsterdam, the Netherlands.

Staff at ACER

Julian Fraillon, *research director*
John Ainley, *project coordinator*
Wolfram Schulz, *assessment coordinator*
Tim Friedman, *project researcher*
Daniel Duckworth, *test development*
Karin Hohlfeld, *test development*
Eveline Gebhardt, *data analyst*
Renee Chow, *data analyst*
Jorge Fallas, *data analyst*
Louise Wenn, *data analyst*
Greg Macaskill, *data analyst*

International Association for the Evaluation of Educational Achievement (IEA)

IEA provides overall support in coordinating ICILS. The IEA Secretariat in Amsterdam, the Netherlands, is responsible for membership, translation verification, and quality control monitoring. The IEA Data Processing and Research Center (DPC) in Hamburg, Germany, is mainly responsible for sampling procedures and processing ICILS data.

Staff at the IEA Secretariat

Dirk Hastedt, *executive director*
Paulína Koršňáková, *director of the IEA Secretariat*
David Ebbs, *research officer (translation verification)*
Alana Yu, *publications officer*
Roel Burgers, *financial manager*
Isabelle Gemin, *financial assistant*

Staff at the IEA Data Processing and Research Center (DPC)

Heiko Sibberns, *director*
Ralph Carstens, *co-project manager*
Michael Jung, *co-project manager*
Sabine Meinck, *researcher (sampling)*
Robert Whitwell, *researcher (sampling)*
Sabine Tieck, *researcher (sampling)*
Diego Cortes, *researcher (sampling)*
Duygu Savasci, *researcher (sampling)*
Dirk Oehler, *research analyst*
Christine Busch, *research analyst*
Tim Daniel, *research analyst*
Sebastian Meyer, *research analyst*
Alena Becker, *research analyst*
Hannah Kohler, *research analyst*
Meng Xue, *head of software unit*
Limiao Duan, *programmer*
Devi Potham Rajendra Prasath, *programmer*
Christian Harries, *programmer*
Poornima Mamadapur, *software tester*
Bettina Wietzorek, *meeting and seminar coordinator*

SoNET Systems

SoNET Systems was responsible for developing the software systems underpinning the computer-based student assessment instruments. This work included development of the test and questionnaire items, the assessment delivery system, and the web-based translation, scoring, and data-management modules.

Staff at SoNET Systems

Mike Janic, *managing director*
Stephen Birchall, *general manager of software development*
Erhan Halil, *senior analyst programmer*
Rakshit Shingala, *analyst programmer*
Stephen Ainley, *quality assurance*
Ranil Weerasinghe, *quality assurance*

ICILS Project Advisory Committee (PAC)

PAC has, from the beginning of the project, advised the international study center and its partner institutions during regular meetings.

PAC members

John Ainley (chair), *ACER, Australia*
Ola Erstad, *University of Oslo, Norway*
Kathleen Scalise, *University of Oregon, United States*
Alfons ten Brummelhuis, *Kennisnet, the Netherlands*

ICILS sampling referee

Jean Dumais from Statistics Canada in Ottawa 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 (NRCs) played a crucial role in the study's development. They provided policy- and content-oriented advice on developing the instruments and were responsible for the implementation of ICILS in the participating countries.

Australia
Lisa DeBortoli
Australian Council for Educational Research (ACER)

Buenos Aires (Argentina)
Silvia Montoya
Assessment and Accountability, Ministry of Education

Canada
Mélanie Labrecque
Council of Ministers of Education (CMEC)

Chile
Gabriela Cares
Education Quality Assurance Agency

Croatia
Michelle Braš Roth
National Center for External Evaluation of Education

Czech Republic
Josef Basl
Czech School Inspectorate

Denmark
Jeppe Bundsgaard
Department of Education, Aarhus University

Germany
Wilfried Bos
Institute for School Development Research, TU Dortmund University

Birgit Eickelmann
Institute for Educational Science, University of Paderborn

Hong Kong SAR
Nancy Law
Centre for Information Technology in Education, the University of Hong Kong

Korea, Republic of
Soojin Kim
Korea Institute for Curriculum and Evaluation

Lithuania
Eugenijus Kurilovas
Asta Buineviciute
Center of Information Technologies in Education

Netherlands
Martina Meelissen
Department of Research Methodology, Measurement, and Data Analysis, University of Twente

Alfons ten Brummelhuis
Kennisnet

Norway
Inger Throndsen
Department of Teacher Education and School Research, University of Oslo

Geir Ottestad
Norwegian Center for ICT in Education

Poland
Kamil Sijko
The Educational Research Institute (IBE)

Russian Federation
Svetlana Avdeeva
National Training Foundation (NTF)

Slovak Republic
Andrea Galádová
National Institute for Certified Educational Measurements (NUCEM)

Slovenia
Eva Klemenčič
Barbara Brecko (field trial)
Center for Applied Epistemology, Educational Research Institute

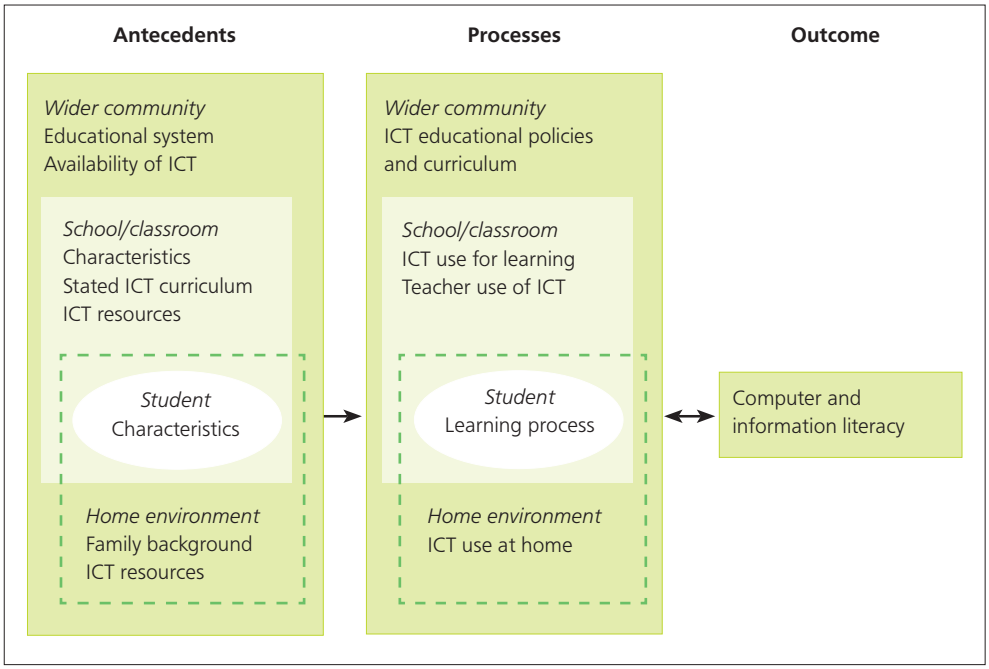
Switzerland
Per Bergamin
Swiss Distance University of Applied Sciences

Thailand
Chaiwuti Lertwanasiriwan
Institute for the Promotion of Teaching Science and Technology (IPST)

Turkey
Gülçin Öz
Meral Alkan (field trial)
Ministry of National Education, General Directorate of Innovation and Educational Technologies

APPENDIX B:
Contexts for CIL learning and learning outcomes

Table B.1: Contexts for CIL learning and learning outcomes



APPENDIX C:

Characteristics of national samples

This appendix describes, for each education system participating in ICILS 2013, the 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 could differ. It was particularly common for a school to count as participating in the student survey but not in the teacher survey; however, the reverse scenario was also possible.

If the school participation status in the two ICILS 2013 surveys differed, the information in the tables in this appendix is displayed in two separate tables. If the status counts were identical in both surveys, the results are displayed in one combined table. Please note also that the numbers of categories for specific stratification variables are sometimes displayed in brackets in the tables.

C.1 Australia

- School-level exclusions consisted of geographically remote schools, distance education schools, schools for children with special needs, alternative curriculum schools, intensive English language schools, schools instructing in a language and curriculum other than English, hospital schools, and correctional schools. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students and nonnative language speakers.
- Explicit stratification was performed by geographic location (remote versus nonremote) and within the nonremote stratum by state/territory, resulting in nine explicit strata.
- Implicit stratification was applied by socioeconomic index for areas (SEIFA) within state/territory (six), sector within state/territory (Catholic, government, independent), geographic location within state/territory (metropolitan, provincial), and state/territory within stratum of remote schools (eight), giving a total of 183 implicit strata.
- The sample was disproportionally allocated to explicit strata. Schools in smaller states or territories were oversampled to obtain more reliable estimates.

Table C.1.1: Allocation of student sample in Australia

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Australian Capital Territory (ACT)	20	0	20	0	0	0
New South Wales (NSW)	46	0	45	0	0	1
Victoria (VIC)	46	0	46	0	0	0
Queensland (QLD)	46	0	45	0	0	1
South Australia (SA)	46	0	42	2	1	1
Western Australia (WA)	46	0	46	0	0	0
Tasmania (TAS)	40	0	38	0	0	2
Northern Territory (NT)	15	1	13	0	0	1
Remote schools	20	0	13	0	0	7
Total	**325**	**1**	**308**	**2**	**1**	**13**

Note: Nine schools were regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.1.2: Allocation of teacher sample in Australia

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Australian Capital Territory (ACT)	20	0	19	0	0	1
New South Wales (NSW)	46	0	42	0	0	4
Victoria (VIC)	46	0	43	0	0	3
Queensland (QLD)	46	0	40	0	0	6
South Australia (SA)	46	0	43	2	1	0
Western Australia (WA)	46	0	43	0	0	3
Tasmania (TAS)	40	0	33	0	0	7
Northern Territory (NT)	15	1	13	0	0	1
Remote schools	20	0	15	0	0	5
Total	**325**	**1**	**291**	**2**	**1**	**30**

Note: Twenty-three schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.2 Chile

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than five students in the target grade), and geographically inaccessible schools. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by overlap with Grade 9 (two levels), school administration type (public, private-subsidized, private), and urbanization (rural, urban), resulting in 12 explicit strata.
- Implicit stratification was applied by national assessment performance group for mathematics (four levels), giving a total of 42 implicit strata.
- The sample was disproportionally allocated to explicit strata. Private and rural schools were oversampled to obtain more reliable estimates for the corresponding subsamples.

Table C.2.1: Allocation of student and teacher sample in Chile

School Participation Status: Student Survey and Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Grades 8 & 9: Private, rural	4	0	4	0	0	0
Grades 8 & 9: Private, urban	46	0	41	4	1	0
Grades 8 & 9: Private-subsidized, rural	4	0	4	0	0	0
Grades 8 & 9: Private-subsidized, urban	36	2	32	2	0	0
Grades 8 & 9: Public, rural	4	0	4	0	0	0
Grades 8 & 9: Public, urban	10	0	9	1	0	0
Grade 8: Private, urban	2	0	1	1	0	0
Grade 8: Private-subsidized, rural	4	0	4	0	0	0
Grade 8: Private-subsidized, urban	18	1	16	1	0	0
Grade 8: Public, rural	14	0	13	0	0	0
Grade 8: Public, urban	38	2	35	1	0	0
Total	**180**	**5**	**163**	**10**	**1**	**0**

Notes:
No schools with a student participation rate below 50% were found.
No schools with a teacher participation rate below 50% were found.

C3 Croatia

- School-level exclusions consisted of schools for students with special educational needs, schools where students were taught in a language different from Croatian, and schools where the curriculum differed in major aspects to the mainstream curriculum. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by region, resulting in six explicit strata.
- Implicit stratification was applied by urbanization level (larger cities, towns, other), giving a total of 16 implicit strata.
- The sample was disproportionally allocated to explicit strata. Schools were oversampled to obtain more reliable estimates for each region.

Table C.3.1: Allocation of student sample in Croatia

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Central Croatia	30	0	30	0	0	0
Eastern Croatia	30	0	29	0	0	1
Northern Croatia	30	0	29	0	0	1
Western Croatia	30	0	27	0	0	3
Southern Croatia	30	0	27	0	0	3
City of Zagreb	30	0	28	0	0	2
Total	**180**	**0**	**170**	**0**	**0**	**10**

Note: Ten schools were regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.3.2: Allocation of teacher sample in Croatia

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Central Croatia	30	0	30	0	0	0
Eastern Croatia	30	0	30	0	0	0
Northern Croatia	30	0	30	0	0	0
Western Croatia	30	0	30	0	0	0
Southern Croatia	30	0	30	0	0	0
City of Zagreb	30	0	29	0	0	1
Total	**180**	**0**	**179**	**0**	**0**	**1**

Note: One school was regarded as nonparticipating because the within-school participation rate was below 50%.

C.4 Czech Republic

- School-level exclusions consisted of schools for children with special educational needs and schools with Polish as the language of instruction. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by school type (elementary schools, multiyear gymnasium), and region (six), resulting in 12 explicit strata.
- Implicit stratification was applied by region (14), giving a total of 28 implicit strata.
- The sample was disproportionally allocated to explicit strata. Multiyear gymnasium schools were oversampled to obtain reliable estimates for group comparisons across school types.

Table C.4.1: Allocation of student and teacher sample in the Czech Republic

School Participation Status: Student Survey and Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Elementary schools: Other regions	84	0	83	1	0	0
Elementary schools: Karlovarsky	4	0	4	0	0	0
Elementary schools: Ustecky	11	0	11	0	0	0
Elementary schools: Liberecky	6	0	6	0	0	0
Elementary schools: Olomoucky	8	0	8	0	0	0
Elementary schools: Moravskoslezsky	17	0	17	0	0	0
Multiyear gymnasium: Other regions	28	0	28	0	0	0
Multiyear gymnasium: Karlovarsky	2	0	2	0	0	0
Multiyear gymnasium: Ustecky	2	0	2	0	0	0
Multiyear gymnasium: Liberecky	2	0	2	0	0	0
Multiyear gymnasium: Olomoucky	2	0	2	0	0	0
Multiyear gymnasium: Moravskoslezsky	4	0	4	0	0	0
Total	**170**	**0**	**169**	**1**	**0**	**0**

Notes:
No schools with a student participation rate below 50% were found.
No schools with a teacher participation rate below 50% were found.

C.5 Denmark

- School-level exclusions consisted of schools for children with special educational needs, day as well as day and night treatment centers, schools with fewer than five students in the target grade, and special schools for youth. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- No explicit stratification was performed.
- Implicit stratification was applied by region (five) and institutional type (*Efterskoler, Folkeskoler, Friskoler/private Grundskoler*), giving a total of 15 implicit strata.
- Participation rates in both the student and the teacher survey were low. Results from both surveys were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.

Table C.5.1: Allocation of student sample in Denmark

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
None	150	9	54	25	24	38
Total	**150**	**9**	**54**	**25**	**24**	**38**

Note: Four schools were regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.5.2: Allocation of teacher sample in Denmark

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
None	150	9	45	19	18	59
Total	**150**	**9**	**45**	**19**	**18**	**59**

Note: Twenty-seven schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.6 Germany

- School-level exclusions consisted of schools for children with special educational needs and Waldorf schools. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by federal state (Berlin, other federal states) and track (gymnasium, nongymnasium, special needs schools), resulting in five explicit strata.
- Implicit stratification was applied by federal state (16 levels) and school type within stratum nongymnasium (four levels), giving a total of 55 implicit strata.
- Participation rates in the teacher survey were low. Results pertaining to the teacher survey were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.

Table C.6.1: Allocation of student sample in Germany

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Berlin: Gymnasiums	2	0	0	1	0	1
Berlin: Nongymnasiums	4	0	2	0	0	2
Gymnasiums (all federal states except Berlin)	54	0	44	8	1	1
Nongymnasiums (all federal states except Berlin)	86	1	60	11	6	8
Special needs schools	4	0	2	1	0	1
Total	**150**	**1**	**108**	**21**	**7**	**13**

Note: Seven schools were regarded as nonparticipating because the within-school participation rate was below 50%. One school was excluded after the main survey but was accounted for in the school-level exclusions. It was set to out-of-scope during weighting to ensure that it was not double-counted in the exclusion rates.

Table C.6.2: Allocation of teacher sample in Germany

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Berlin: Gymnasiums	2	0	0	1	1	0
Berlin: Nongymnasiums	4	0	1	0	0	3
Gymnasiums (all federal states except Berlin)	54	0	40	5	0	9
Nongymnasiums (all federal states except Berlin)	86	1	56	8	6	15
Special needs schools	4	0	2	1	0	1
Total	**150**	**1**	**99**	**15**	**7**	**28**

Note: Twenty-two schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.7 Hong Kong SAR

- School-level exclusions consisted of schools for children with special needs, private schools, and international schools. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers. The overall exclusion rate exceeded the ICILS 2013 threshold for exclusions (6.5%).
- Explicit stratification was performed by areas of similar median monthly income, resulting in four explicit strata.
- Implicit stratification was applied by gender (coeducation, boys, girls) and finance type (three), giving a total of 29 implicit strata.
- The sample was disproportionally allocated to explicit strata. Oversampling of schools was done to obtain reliable estimates for group comparisons across different income brackets.
- Participation rates in both the student and the teacher survey were low. Results from both surveys were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.

Table C.7.1: Allocation of student sample in Hong Kong SAR

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Monthly income ≤15,000 HK$	50	0	33	3	0	14
Monthly income 15,001 to 18,500 HK$	30	0	25	1	0	4
Monthly income 18,501 to 20,000 HK$	30	0	21	1	0	8
Monthly income >20,000 HK$	40	0	32	2	0	6
Total	**150**	**0**	**111**	**7**	**0**	**32**

Note: Five schools were regarded as nonparticipating because the within-school participation rate was below 50%. Data from one school were rejected because of incorrect student lists.

Table C.7.2: Allocation of teacher sample in Hong Kong SAR

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Monthly income ≤15,000 HK$	50	0	29	3	0	18
Monthly income 15,000 to 18,500 HK$	30	0	20	1	0	9
Monthly income 18,501 to 20,000 HK$	30	0	21	1	0	8
Monthly income >20,000 HK$	40	0	30	2	0	8
Total	**150**	**0**	**100**	**7**	**0**	**43**

Note: Seventeen schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.8 Korea, Republic of

- School-level exclusions consisted of geographically inaccessible schools, schools with fewer than five students in the target grade, and schools with different curriculums (physical-education middle schools). Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by regions, resulting in 16 explicit strata.
- Implicit stratification was applied by gender (boys, girls, mixed), giving a total of 48 implicit strata.

Table C.8.1: Allocation of student and teacher sample in the Republic of Korea

School Participation Status: Student Survey and Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Seoul	26	0	26	0	0	0
Pusan	10	0	10	0	0	0
Taegu	8	0	8	0	0	0
Inchon	8	0	8	0	0	0
Kwangju	6	0	6	0	0	0
Taejon	5	0	5	0	0	0
Ulsan	4	0	4	0	0	0
Kyunggi-do	37	0	37	0	0	0
Kangwon-do	4	0	4	0	0	0
Chungchongbuk-do	5	0	5	0	0	0
Chungchongnam-do	6	0	6	0	0	0
Chollabuk-do	6	0	6	0	0	0
Chollanam-do	5	0	5	0	0	0
Kyongsangbuk-do	8	0	8	0	0	0
Kyongsangnam-do	10	0	10	0	0	0
Cheju-do	2	0	2	0	0	0
Total	**150**	**0**	**150**	**0**	**0**	**0**

Notes:
No schools with a student participation rate below 50% were found.
No schools with a teacher participation rate below 50% were found.

C.9 Lithuania

- School-level exclusions consisted of schools for children with special needs, Lithuanian public schools with fewer than seven students, and schools located in prisons. Within-school exclusions consisted of intellectually-disabled students and functionally-disabled students.
- Explicit stratification was performed by teaching language (Lithuanian, minority language), organizational type (public, private), and urbanization (urban, rural), resulting in seven explicit strata.
- Implicit stratification was applied by language of instruction (eight), giving a total of 16 implicit strata.
- The sample was disproportionally allocated to explicit strata. Private schools and minority language schools were oversampled to obtain reliable estimates for group comparisons.

Table C.9.1: Allocation of student sample in Lithuania

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Lithuanian: Private, rural	3	0	3	0	0	0
Lithuanian: Private, urban	12	1	7	0	0	4
Lithuanian: Public, rural	90	3	79	6	0	2
Lithuanian: Public, urban	46	1	43	2	0	0
Minority language: Public, rural	12	0	11	1	0	0
Minority language: Public, urban	16	0	10	0	0	6
Total	**179**	**5**	**153**	**9**	**0**	**12**

Note: One school was regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.9.2: Allocation of teacher sample in Lithuania

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Lithuanian: Private, rural	3	0	3	0	0	0
Lithuanian: Private, urban	12	1	7	0	0	4
Lithuanian: Public, rural	90	3	80	6	0	1
Lithuanian: Public, urban	46	1	43	2	0	0
Minority language: Public, rural	12	0	11	1	0	0
Minority language: Public, urban	16	0	10	0	0	6
Total	**179**	**5**	**154**	**9**	**0**	**11**

Note: No schools with a teacher participation rate below 50% were found.

C.10 Netherlands

- School-level exclusions consisted of schools for children with special educational needs, very small schools (fewer than 10 students in the target grade), and international schools with English as the language of instruction. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed for tracks at schools (VMBO/PRO, HAVO/VWO, mixed), resulting in three explicit strata.
- No implicit stratification was applied.
- A deviation in the student sampling procedure for the student survey was approved as follows: class sampling was performed in 27 schools, and one intact classroom was sampled in each of these schools.
- Participation rates in both the student and the teacher survey were low. Results from both surveys were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.

Table C.10.1: Allocation of student sample in the Netherlands

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
PRO/VMBO: Practical training and prevocational secondary education	44	0	25	5	9	5
HAVO/VWO: Senior general secondary education and preuniversity education	27	0	13	2	5	7
PRO/VMBO/HAVO/VWO: Practical training, prevocational secondary education, senior general secondary education, and preuniversity education	79	1	36	20	6	15
Total	**150**	**1**	**74**	**27**	**20**	**27**

Note: Three schools were regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.10.2: Allocation of teacher sample in the Netherlands

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
PRO/VMBO: Practical training and prevocational secondary education	44	0	20	4	7	13
HAVO/VWO: Senior general secondary education and preuniversity education	27	0	12	2	2	11
PRO/VMBO/HAVO/VWO: Practical training, prevocational secondary education, senior general secondary education, and preuniversity education	79	1	26	19	4	28
Total	**150**	**1**	**58**	**25**	**13**	**52**

Note: Twenty-eight schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.11 Norway (Grade 9)

- School-level exclusions consisted of schools for children with special educational needs, Steiner schools, schools with fewer than five students in the target grade, international schools with another language of instruction, and schools with Saami as the language of instruction. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers. The overall exclusion rate exceeded the ICILS 2013 threshold for exclusions (6.1%).
- Explicit stratification was implemented by performance groups, resulting in four explicit strata.
- Implicit stratification was applied by language (Bokmål, Nynorsk), giving a total of eight implicit strata.
- Participation rates in the teacher survey were low. Results pertaining to the teacher survey were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.
- Because Norway decided to survey students and their teachers at the end of their ninth grade instead of assessing students at Grade 8, their results were annotated accordingly in the reporting tables in the ICILS 2013 international report.

Table C.11.1: Allocation of student sample in Norway

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Low performance	40	0	33	4	0	3
Medium performance	60	0	51	3	1	5
High performance	46	1	40	3	0	2
Performance unknown	4	0	3	0	0	1
Total	**150**	**1**	**127**	**10**	**1**	**11**

Note: No schools with student participation rate below 50% were found.

Table C.11.2: Allocation of teacher sample in Norway

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Low performance	40	0	27	2	0	11
Medium performance	60	0	43	3	1	13
High performance	46	1	34	3	0	8
Performance unknown	4	0	3	0	0	1
Total	**150**	**1**	**107**	**8**	**1**	**33**

Note: Twenty-two schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.12 Poland

- School-level exclusions consisted of schools for children with special needs, schools with fewer than nine students in the target grade, and other special schools. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by school type (public, private), performance level (three), and creative schools, resulting in seven explicit strata.
- Implicit stratification was applied by urbanization within the public stratum (village, small city, medium city, large city), giving a total of 16 implicit strata.
- The sample was disproportionally allocated to explicit strata, and all "creative schools" were included in the survey to obtain more reliable estimates for group comparisons.

Table C.12.1: Allocation of student sample in Poland

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Creative schools	12	0	12	0	0	0
Regular schools: Low score, public	30	0	22	5	2	1
Regular schools: Low score, private	2	0	1	0	0	1
Regular schools: Medium score, public	80	0	70	10	0	0
Regular schools: Medium score, private	2	0	1	1	0	0
Regular schools: High score, public	30	0	26	4	0	0
Regular schools: High score, private	2	0	2	0	0	0
Total	**158**	**0**	**134**	**20**	**2**	**2**

Note: One school was regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.12.2: Allocation of teacher sample in Poland

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Creative schools	12	0	12	0	0	0
Regular schools: Low score, public	30	0	22	5	2	1
Regular schools: Low score, private	2	0	2	0	0	0
Regular schools: Medium score, public	80	0	70	10	0	0
Regular schools: Medium score, private	2	0	1	1	0	0
Regular schools: High score, public	30	0	26	4	0	0
Regular schools: High score, private	2	0	2	0	0	0
Total	**158**	**0**	**135**	**20**	**2**	**1**

Note: No schools with a teacher participation rate below 50% were found.

C.13 Russian Federation

- A sample of 43 regions out of 83 was first sampled with PPS. The largest 14 regions were sampled with certainty. A sample of schools was then drawn within each sampled region. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers. The overall exclusion rate exceeded the ICILS 2013 threshold for exclusions (5.9%).
- School-level exclusions consisted of schools for children with special educational needs, schools with fewer than four students in the target grade, and evening schools.
- Explicit stratification for the second sampling stage was performed according to regions, resulting in 43 explicit strata. Small school samples within regions necessitated disproportional sample allocations.
- Implicit stratification was applied by urbanization (rural, urban), giving a total of 84 implicit strata.
- Students were tested at the beginning of Grade 9 rather than at the end of Grade 8 (about seven months after the regular testing time). Students referred to their current school year when answering the student background questionnaire. Because of this delayed survey administration, teachers filled in their questionnaires retrospectively and referred to the previous school year when they were teaching Grade 8 students. Student and teacher sampling occurred before the summer break in order to maintain compliance with the international target population definitions; three percent of the sampled students and five percent of the sampled teachers had left the school permanently before survey administration.

Table C.13.1: Allocation of student sample in the Russian Federation

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Pskov obl*	4	0	4	0	0	0
Sankt Petersburg*	6	0	6	0	0	0
Moscow*	12	0	12	0	0	0
Moscow obl*	10	0	10	0	0	0
Samara obl*	6	0	6	0	0	0
N Novgorod obl*	6	0	6	0	0	0
Tatarstan*	6	0	6	0	0	0
Bashkortostan*	6	0	6	0	0	0
Krasnodar kr*	6	0	6	0	0	0
Dagestan*	6	0	6	0	0	0
Rostov obl*	6	0	5	0	0	1
Chelyabinsk obl*	6	0	6	0	0	0
Sverdlovsk obl*	6	0	6	0	0	0
Krasnoyarsk kr*	6	0	6	0	0	0
Novgorod obl	4	0	4	0	0	0
Kaliningrad obl	4	0	4	0	0	0
Vologda obl	4	0	4	0	0	0
Voronezh obl	4	0	4	0	0	0
Tula obl	4	0	4	0	0	0
Bransk obl	4	0	4	0	0	0
Kursk obl	4	0	4	0	0	0
Razan obl	4	0	4	0	0	0
Kaluga obl	4	0	4	0	0	0
Kostroma obl	4	0	4	0	0	0
Ulianovsk obl	4	0	3	0	0	1
Chuvashia	4	0	4	0	0	0
Orenburg obl	4	0	4	0	0	0
Saratov obl	4	0	4	0	0	0
Perm kr	4	0	4	0	0	0
Stavropol kr	4	0	4	0	0	0
Volgograd obl	4	0	4	0	0	0
Astrakhan obl	4	0	4	0	0	0
Alania	4	0	4	0	0	0
Iamal-Nenets ok	4	0	4	0	0	0
Hanty-Mansii ok	4	0	4	0	0	0
Irkutsk obl	4	0	4	0	0	0
Kemerovo obl	4	0	4	0	0	0
Novosibirsk obl	4	0	4	0	0	0
Altay kr	4	0	4	0	0	0
Zabaykalski kr	4	0	4	0	0	0
Tomsk obl	4	0	4	0	0	0
Amur obl	4	0	4	0	0	0
Sakha	4	0	4	0	0	0
Total	**208**	**0**	**206**	**0**	**0**	**2**

Notes:

*Certainty regions.

One school was regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.13.2: Allocation of teacher sample in the Russian Federation

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Pskov obl*	4	0	3	0	0	1
Sankt Petersburg*	6	0	6	0	0	0
Moscow*	12	0	12	0	0	0
Moscow obl*	10	0	10	0	0	0
Samara obl*	6	0	6	0	0	0
N Novgorod obl*	6	0	6	0	0	0
Tatarstan*	6	0	6	0	0	0
Bashkortostan*	6	0	6	0	0	0
Krasnodar kr*	6	0	6	0	0	0
Dagestan*	6	0	6	0	0	0
Rostov obl*	6	0	6	0	0	0
Chelyabinsk obl*	6	0	6	0	0	0
Sverdlovsk obl*	6	0	6	0	0	0
Krasnoyarsk kr*	6	0	6	0	0	0
Novgorod obl	4	0	4	0	0	0
Kaliningrad obl	4	0	4	0	0	0
Vologda obl	4	0	4	0	0	0
Voronezh obl	4	0	4	0	0	0
Tula obl	4	0	4	0	0	0
Bransk obl	4	0	4	0	0	0
Kursk obl	4	0	4	0	0	0
Razan obl	4	0	4	0	0	0
Kaluga obl	4	0	4	0	0	0
Kostroma obl	4	0	4	0	0	0
Ulianovsk obl	4	0	4	0	0	0
Chuvashia	4	0	4	0	0	0
Orenburg obl	4	0	4	0	0	0
Saratov obl	4	0	4	0	0	0
Perm kr	4	0	4	0	0	0
Stavropol kr	4	0	4	0	0	0
Volgograd obl	4	0	4	0	0	0
Astrakhan obl	4	0	4	0	0	0
Alania	4	0	4	0	0	0
Iamal-Nenets ok	4	0	4	0	0	0
Hanty-Mansii ok	4	0	4	0	0	0
Irkutsk obl	4	0	4	0	0	0
Kemerovo obl	4	0	4	0	0	0
Novosibirsk obl	4	0	4	0	0	0
Altay kr	4	0	4	0	0	0
Zabaykalski kr	4	0	4	0	0	0
Tomsk obl	4	0	4	0	0	0
Amur obl	4	0	4	0	0	0
Sakha	4	0	4	0	0	0
Total	**208**	**0**	**207**	**0**	**0**	**1**

Notes:

*Certainty regions.

One school was regarded as nonparticipating because the within-school participation rate was below 50%.

C.14 Slovak Republic

- School-level exclusions consisted of schools in the system of education for children with special educational needs, schools with fewer than five students in the target grade, and schools with another language of instruction. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by school type (grammar school, gymnasium) and language (Slovakian, Hungarian), resulting in four explicit strata.
- Implicit stratification was applied by region (eight), giving a total of 24 implicit strata.
- The sample was disproportionally allocated to explicit strata. Gymnasiums and Hungarian schools were oversampled to obtain more reliable estimates for comparisons between school types and between language groups.

Table C.14.1: Allocation of student and teacher sample in the Slovak Republic

School Participation Status: Student Survey and Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Grammar: Hungarian	28	0	25	1	0	2
Grammar: Slovakian	116	1	110	4	0	0
Gymnasium: Hungarian	2	0	2	0	0	0
Gymnasium: Slovakian	28	3	19	6	0	0
Total	**174**	**4**	**156**	**11**	**0**	**2**

Notes:
No schools with a student participation rate below 50% were found.
No schools with a teacher participation rate below 50% were found.

C.15 Slovenia

- School-level exclusions consisted of schools for children with special needs, Italian schools, and Waldorf schools. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by region, resulting in 12 explicit strata.
- Implicit stratification was applied by performance levels (three), giving a total of 36 implicit strata.
- The sample was disproportionally allocated to explicit strata. Small regions were oversampled to obtain more reliable estimates for comparisons across regions.

Table C.15.1: Allocation of student sample in Slovenia

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Pomurska	20	0	20	0	0	0
Podravska	20	0	19	0	0	1
Koroška	17	0	16	0	0	1
Savinjska	20	0	18	2	0	0
Zasavska	7	0	7	0	0	0
Spodnjeposavska	22	0	21	0	0	1
Jugovzhodna Slovenija	20	0	19	1	0	0
Osrednjeslovenska	20	0	16	4	0	0
Gorenjska	20	0	19	1	0	0
Notranjsko-Kraška	16	0	16	0	0	0
Goriška	20	0	17	2	0	1
Obalno-Kraška	21	0	20	0	0	1
Total	**223**	**0**	**208**	**10**	**0**	**5**

Note: One school was regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.15.2: Allocation of teacher sample in Slovenia

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Pomurska	20	0	20	0	0	0
Podravska	20	0	20	0	0	0
Koroška	17	0	16	0	0	1
Savinjska	20	0	18	2	0	0
Zasavska	7	0	7	0	0	0
Spodnjeposavska	22	0	21	0	0	1
Jugovzhodna Slovenija	20	0	19	1	0	0
Osrednjeslovenska	20	0	14	3	0	3
Gorenjska	20	0	19	1	0	0
Notranjsko-Kraška	16	0	16	0	0	0
Goriška	20	0	17	2	0	1
Obalno-Kraška	21	0	18	0	0	3
Total	**223**	**0**	**205**	**9**	**0**	**9**

Note: Five schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.16 Switzerland

- School-level exclusions consisted of schools in the system of education for children with special needs and schools with fewer than five students in the target grade. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by cantons and regions, resulting in eight explicit strata.
- Implicit stratification was applied by organization (public, private) and language (German, French, Italian), giving a total of 14 implicit strata.
- The sample was disproportionally allocated to explicit strata. The cantons Tessin and Valais were oversampled to obtain more reliable population estimates for these cantons.
- A deviation in the student sampling procedure for the student survey was approved as follows: class sampling was performed for nine schools, and one to three classrooms were sampled in the selected schools.
- Participation rates in the student survey were low. Results pertaining to the student survey were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.
- Participation rates in the teacher survey were particularly low. The Swiss results for this survey were therefore not included in the ICILS 2013 international report and its data are not included in the ICILS database.

Table C.16.1: Allocation of student sample in Switzerland

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Canton Bejune (French)	4	0	2	2	0	0
Canton Jura	4	0	3	0	1	0
Canton Neuchâtel	7	0	6	0	0	1
Canton Tessin	20	0	20	0	0	0
Other cantons	120	3	28	11	11	67
Valais Region 1	5	0	5	0	0	0
Valais Region 2	5	0	4	0	0	1
Valais Region 3	5	0	5	0	0	0
Total	**170**	**3**	**73**	**13**	**12**	**69**

Note: No schools with a student participation rate below 50% were found.

Table C.16.2: Allocation of teacher sample in Switzerland

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Canton Bejune (French)	4	0	1	2	0	1
Canton Jura	4	0	0	0	1	3
Canton Neuchâtel	7	0	2	0	0	5
Canton Tessin	20	0	17	0	0	3
Other cantons	120	3	20	7	11	79
Valais Region 1	5	0	4	0	0	1
Valais Region 2	5	0	4	0	0	1
Valais Region 3	5	0	5	0	0	0
Total	**170**	**3**	**53**	**9**	**12**	**93**

Note: Twenty-one schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.17 Thailand

- School-level exclusions consisted of schools for children with special educational needs and schools with fewer than six students in the target grade. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by jurisdiction, resulting in five explicit strata.
- Implicit stratification was applied by region (five), giving a total of 20 implicit strata.
- The sample was disproportionally allocated to explicit strata. Small jurisdictions were oversampled to obtain more reliable estimates for comparisons between jurisdictions.

Table C.17.1: Allocation of student sample in Thailand

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Basic Education Commission (OBEC)	94	0	84	4	1	5
Private Education Commission (OPEC)	30	0	25	3	0	2
Bangkok Metropolitan Administration (BMA)	30	0	26	3	1	0
Department of Local Administration (DLA)	30	0	28	2	0	0
Higher Education Commission (OHEC)	26	1	21	0	0	4
Total	**210**	**1**	**184**	**12**	**2**	**11**

Note: One school was regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.17.2: Allocation of teacher sample in Thailand

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Basic Education Commission (OBEC)	94	0	78	3	2	11
Private Education Commission (OPEC)	30	0	25	3	0	2
Bangkok Metropolitan Administration (BMA)	30	0	24	3	1	2
Department of Local Administration (DLA)	30	0	25	2	0	3
Higher Education Commission (OHEC)	26	1	18	0	0	7
Total	**210**	**1**	**170**	**11**	**3**	**25**

Note: Fifteen schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.18 Turkey

- School-level exclusions consisted of schools for children with special educational needs, geographically inaccessible schools, schools with fewer than six students in the target grade, private foreign schools, and music and ballet schools. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by organization (public, private), resulting in two explicit strata.
- Implicit stratification was applied by geographical regions (seven) within the public school stratum, giving a total of eight implicit strata.

Table C.18.1: Allocation of student sample in Turkey

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Public	144	0	134	1	0	9
Private	6	0	6	0	0	0
Total	**150**	**0**	**140**	**1**	**0**	**9**

Note: One school was regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.18.2: Allocation of teacher sample in Turkey

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Public	144	0	143	1	0	0
Private	6	0	6	0	0	0
Total	**150**	**0**	**149**	**1**	**0**	**0**

Note: No schools with a teacher participation rate below 50% were found.

Benchmarking Participants

C.19 City of Buenos Aires, Argentina

- School-level exclusions consisted of schools for children with special educational needs. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was implemented by organization (public, private), resulting in two explicit strata.
- Implicit stratification was applied by socioeconomic area index (three), giving a total of six implicit strata.
- Participation rates in the student survey were low. Results pertaining to the student survey were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.
- Participation rates in the teacher survey were particularly low. Results pertaining to the teacher survey were therefore not included in the ICILS 2013 international report and its data are not included in the ICILS database.

Table C.19.1: Allocation of student sample in the City of Buenos Aires, Argentina

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Public	50	0	29	0	0	21
Private	0	0	39	0	0	11
Total	**50**	**0**	**68**	**0**	**0**	**32**

Note: Eight schools with a student participation rate below 50% were found.

Table C.19.2: Allocation of teacher sample in the City of Buenos Aires, Argentina

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
Public	50	0	20	0	0	30
Private	50	0	29	0	0	21
Total	**100**	**0**	**49**	**0**	**0**	**51**

Note: Twenty-seven schools with a student participation rate below 50% were found.

C.20 Newfoundland and Labrador, Canada

- School-level exclusions consisted of schools with native languages of instruction and schools with fewer than six students in the target grade. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers. The overall exclusion rate exceeded the ICILS 2013 threshold for exclusions (7.6%).
- Explicit stratification was performed by language (English/French), resulting in two explicit strata.
- No implicit stratification was applied.
- A deviation in the standard sampling procedure for the teacher survey was approved as follows: only five teachers were sampled per school. Given the small school sizes, this procedure resulted in a census of eligible teachers in about 33 percent of all sampled schools in Newfoundland and Labrador.

Table C.20.1: Allocation of student sample in Newfoundland and Labrador, Canada

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
English	151	1	117	0	0	2
French	4	0	1	0	0	0
Total	**155**	**4**	**118**	**0**	**0**	**2**

Note: One school was regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.20.2: Allocation of teacher sample in Newfoundland and Labrador, Canada

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
English	151	1	102	0	0	31
French	4	0	1	0	0	3
Total	**155**	**1**	**103**	**0**	**0**	**34**

Note: Sixteen schools were regarded as nonparticipating because the within-school participation rate was below 50%.

C.21 Ontario, Canada

- School-level exclusions consisted of schools with fewer than seven students. Within-school exclusions consisted of intellectually-disabled students, functionally-disabled students, and nonnative language speakers.
- Explicit stratification was performed by language (English, French, mixed), resulting in three explicit strata.
- Implicit stratification was applied by funding (three) and region (six), giving a total of 27 implicit strata.
- The sample was disproportionally allocated to explicit strata. French schools were oversampled to accommodate better estimates with respect to languages.
- A deviation in the standard sampling procedure for the teacher survey was approved as follows: only five teachers were sampled per school. Given the small school sizes, this procedure resulted in a census of eligible teachers in about 75 percent of all sampled schools in Ontario.
- Participation rates in the teacher survey were low. The results pertaining to the teacher survey were therefore presented in separate sections of the reporting tables in the ICILS 2013 international report.

Table C.21.1: Allocation of student sample in Ontario, Canada

School Participation Status: Student Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
English	120	1	112	2	1	4
French	80	2	77	0	0	1
English and French	2	0	1	0	0	1
Total	**202**	**3**	**190**	**2**	**1**	**6**

Note: Two schools were regarded as nonparticipating because the within-school participation rate was below 50%.

Table C.21.2: Allocation of teacher sample in Ontario, Canada

School Participation Status: Teacher Survey

Explicit Strata	Total Sampled Schools	Ineligible Schools	Participating Schools			Nonparticipating Schools
			Sampled	First replacement	Second replacement	
English	120	1	90	1	1	27
French	80	2	61	0	0	17
English and French	2	0	0	0	0	0
Total	**202**	**3**	**151**	**1**	**1**	**44**

Note: Forty-one schools were regarded as nonparticipating because the within-school participation rate was below 50%.

APPENDIX D:

Items excluded from scaling and student-level variables used for conditioning

Table D.1: Items excluded from scaling

Items at international level			
Country	**Item**	**Observation**	**Item treatment**
All countries	A10K	Unsatisfactory scaling properties	Deleted from database
All countries	B07A	Unsatisfactory scaling properties	Deleted from database
All countries	H04A	Unsatisfactory scaling properties	Deleted from database
All countries	S02Z	Unsatisfactory scaling properties	Deleted from database
All countries	S05Z	Unsatisfactory scaling properties	Deleted from database
All countries	A04Z	Large differential item functioning for gender	Deleted from database
All countries	H07F	Large item-by-country interaction across countries	Deleted from database
All countries	A10BF	Items A10B and A10F showed local dependence	Items merged into A10BF
Items at national level			
Country	**Item**	**Observation**	**Item treatment**
Australia	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Australia	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Australia	B09C	Low scorer reliability	Excluded from scaling (retained in database)
Australia	H07B	Low scorer reliability	Excluded from scaling (retained in database)
Australia	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Australia	H07D	Low scorer reliability	Excluded from scaling (retained in database)
Australia	S08F	Low scorer reliability	Excluded from scaling (retained in database)
Chile	H07D	Scoring error	Deleted from database
Chile	S04A	Scoring error	Deleted from database
Croatia	H07B	Low scorer reliability	Excluded from scaling (retained in database)
Croatia	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Czech Republic	B09D	Large item-by-country interaction	Excluded from scaling (retained in database)
Czech Republic	H07E	Large item-by-country interaction	Excluded from scaling (retained in database)
Germany	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Germany	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Germany	H07B	Low scorer reliability	Excluded from scaling (retained in database)
Germany	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Germany	S06Z	Scoring error	Deleted from database
Germany	S08F	Low scorer reliability	Excluded from scaling (retained in database)

Table D.1: Items excluded from scaling (contd.)

Items at national level			
Country	Item	Observation	Item treatment
Hong Kong SAR	A10G	Large item-by-country interaction	Excluded from scaling (retained in database)
Hong Kong SAR	A10I	Large item-by-country interaction	Excluded from scaling (retained in database)
Hong Kong SAR	H07J	Large item-by-country interaction	Excluded from scaling (retained in database)
Korea, Republic of	H07A	Large item-by-country interaction	Excluded from scaling (retained in database)
Korea, Republic of	H07B	Scoring error	Deleted from database
Lithuania	A10E	Large item-by-country interaction	Excluded from scaling (retained in database)
Lithuania	B09C	Large item-by-country interaction	Excluded from scaling (retained in database)
Lithuania	H07A	Large item-by-country interaction	Excluded from scaling (retained in database)
Netherlands	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	A10I	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	B09C	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	H07B	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	H07G	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	H07J	Low scorer reliability	Excluded from scaling (retained in database)
Netherlands	S08F	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	A10H	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	B09C	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	H07B	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	H07D	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	H07G	Low scorer reliability	Excluded from scaling (retained in database)
Newfoundland and Labrador, Canada	S08F	Low scorer reliability	Excluded from scaling (retained in database)
Norway	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Norway	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Norway	H07D	Low scorer reliability	Excluded from scaling (retained in database)

Table D.1: Items excluded from scaling (contd.)

Items at national level			
Country	Item	Observation	Item treatment
Ontario, Canada	A08Z	Technical failure (French-language instrument)	Deleted from database
Ontario, Canada	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	A10H	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	B09C	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	H07B	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	H07D	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	H07G	Low scorer reliability	Excluded from scaling (retained in database)
Ontario, Canada	S08F	Low scorer reliability	Excluded from scaling (retained in database)
Poland	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Poland	A10E	Large item-by-country interaction	Excluded from scaling (retained in database)
Poland	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Poland	A10I	Large item-by-country interaction	Excluded from scaling (retained in database)
Poland	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	A10A	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	A10C	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	A10E	Large item-by-country interaction	Excluded from scaling (retained in database)
Russian Federation	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	A10H	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	B09C	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	B09E	Large item-by-country interaction	Excluded from scaling (retained in database)
Russian Federation	H07B	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	H07C	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	H07D	Low scorer reliability	Excluded from scaling (retained in database)
Russian Federation	S08F	Low scorer reliability	Excluded from scaling (retained in database)
Slovenia	A10I	Large item-by-country interaction	Excluded from scaling (retained in database)
Slovenia	S01Z	Technical failure	Deleted from database
Switzerland	A08Z	Technical failure (French language instrument)	Deleted from database
Switzerland	A10G	Low scorer reliability	Excluded from scaling (retained in database)
Switzerland	H07C	Low scorer reliability	Excluded from scaling (retained in database)

Table D.1: Items excluded from scaling (contd.)

Items at national level			
Country	**Item**	**Observation**	**Item treatment**
Switzerland	H07D	Low scorer reliability	Excluded from scaling (retained in database)
Thailand	A10E	Low scorer reliability	Excluded from scaling (retained in database)
Thailand	H07D	Low scorer reliability	Excluded from scaling (retained in database)
Thailand	H07J	Low scorer reliability	Excluded from scaling (retained in database)
Turkey	A10E	Large item-by-country interaction	Excluded from scaling (retained in database)
Turkey	B09E	Large item-by-country interaction	Excluded from scaling (retained in database)
Turkey	B09F	Large item-by-country interaction	Excluded from scaling (retained in database)
Turkey	H01Z	Large item-by-country interaction	Excluded from scaling (retained in database)
Turkey	S08B	Large item-by-country interaction	Excluded from scaling (retained in database)
Turkey	S08E	Translation error	Deleted from database

Table D.2: 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
Gender	GENDER	Female Male Missing	10 00 01	Direct
Age	AGE	Value Missing	Copy,0 Mean,1	Direct
Student: Levels of education he/she expected to complete	SISCED (IS1G03)	4. [ISCED Level 5A or 6] 3. [ISCED Level 4 or 5B] 2. [ISCED Level 3] 1. [ISCED Level 2] 0. I do not expect to complete [ISCED Level 2] Missing	Copy,0 Mode,1	Direct
Highest level of parental education	HISCED	4. [ISCED Level 5A or 6] 3. [ISCED Level 4 or 5B] 2. [ISCED Level 3] 1. [ISCED Level 2] 0. I do not expect to complete [ISCED Level 2] Missing	Copy,0 Mode,1	Direct
Country of birth: You Country of birth: Mother or [female guardian] Country of birth: Father or [male guardian]	IS1G04A IS1G04B IS1G04C	1. = [Country of test] 2. = [Other country/Group A] 3. = [Other country/Group B] 4. = [Another country] Missing	Two dummy variables per question, with the national mode as reference category	PCA
Language at home	IS1G05	1 = [Language of test] 2 = [Other language 1] 3 = [Other language 2] 4 = [Another language] Missing	Two dummy variables per question, with the national mode as reference category	PCA

Table D.2: Student background variables used for conditioning (contd.)

Variable	Name	Values	Coding	Regressor
Books at home	IS1G12	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	00000 10000 01000 00100 00010 00001	PCA
Computers at home: Desktop Computers at home: Portable	IS1G13A IS1G13B	Number of computers Missing	Copy,0 Mode,1	PCA
Internet connection at home	IS1G14	1. None 2. Dial-up 3. Broadband (for example, [cable], [DSL], [satellite]) 4. Connection through mobile phone network 5. I know we have internet but I don't know what type of connection it is. Missing	00000 10000 01000 00100 00010 00001	PCA
How long using computer	IS1G15	0. Less than one year 1. At least once a 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 4. Seven years or more Missing	Copy,0 Mode,1	PCA
Computer operating system at home	IS1G16A	1. Windows (PC) 2. Mac OS 3. Other 4. I don't know 5. I do not use a computer at this location Missing	00000 10000 01000 00100 00010 00001	PCA
Computer operating system at school	IS1G16B	1. Windows (PC) 2. Mac OS 3. Other 4. I don't know 5. I do not use a computer at this location Missing	00000 10000 01000 00100 00010 00001	PCA

Table D.2: Student background variables used for conditioning (contd.)

Variable	Name	Values	Coding	Regressor
How often using computer: At home How often using computer: At school How often using computer: At other places (for example, local library, internet cafe)	IS1G17A IS1G17B IS1G17C	0. Never 1. Less than once a month 2. At least once a month but not every week 3. At least once a week but not every day 4. Every day Missing	Copy,0 Mode,1	PCA
Using computer outside school: Creating or editing documents (for example, to write stories or assignments) Using computer outside school: Using a spreadsheet to do calculations, store data, or plot graphs (for example, using [Microsoft EXCEL ®]) Using computer outside school: Creating a simple "slideshow" presentation (for example, using [Microsoft PowerPoint ®]) Using computer outside school: Creating a multimedia presentation (with sound, pictures, video) Using computer outside school: Using education software that is designed to help with your school study (for example, mathematics or reading software) Using computer outside school: Writing computer programs, macros, or scripts (for example, using [Logo, Basic or HTML]) Using computer outside school: Using drawing, painting or graphics software	IS1G18A IS1G18B IS1G18C IS1G18D IS1G18E IS1G18F IS1G18G	0. Never 1. Less than once a month 2. At least once a month but not every week 3. At least once a week but not every day 4. Every day Missing	Copy,0 Mode,1	PCA

Table D.2: Student background variables used for conditioning (contd.)

Variable	Name	Values	Coding	Regressor
Using internet outside school: Searching for information for study or school work	IS1G19A	0. Never 1. Less than once a month 2. At least once a month but not every week 3. At least once a week but not every day 4. Every day Missing	Copy,0 Mode,1	
Using internet outside school: Accessing wikis or online encyclopedia for study or school work	IS1G19B			
Using internet outside school: Communicating with others and using messaging or social networks (for example, instant messaging or [status updates])	IS1G19C			
Using internet outside school: Posting comments to online profiles or blogs	IS1G19D			
Using internet outside school: Asking questions on forums or [question and answer] websites	IS1G19E			PCA
Using internet outside school: Answering other people's questions on forums or websites	IS1G19F			
Using internet outside school: Writing posts for your own blog	IS1G19G			
Using internet outside school: Uploading images or video to an [online profile] or [online community] (for example, Facebook or YouTube)	IS1G19H			
Using internet outside school: Using voice chat (for example, Skype) to chat with friends or family online	IS1G19I			
Using internet outside school: Building or editing a webpage	IS1G19J			
Out-of-school activities: Accessing the internet to find out about places to go or activities to do	IS1G20A	0. Never 1. Less than once a month 2. At least once a month but not every week 3. At least once a week but not every day 4. Every day Missing	Copy,0 Mode,1	
Out-of-school activities: Reading reviews on the internet of things you might want to buy	IS1G20B			
Out-of-school activities: Playing games	IS1G20C			PCA
Out-of-school activities: Listening to music	IS1G20D			
Out-of-school activities: Watching downloaded or streamed video (for example, movies, TV shows or clips)	IS1G20E			
Out-of-school activities: Using the internet to get news about things I am interested in	IS1G20F			

TableD.2: Student background variables used for conditioning (contd.)

Variable	Name	Values	Coding	Regressor
Computer use: Preparing reports or essays	IS1G21A	0. Never		
Computer use: Preparing presentations	IS1G21B	1. Less than once a month	Copy,0	
Computer use: Working with other students from your own school	IS1G21C	2. At least once a month but not every week 3. At least once a week but not every day		PCA
Computer use: Working with other students from other schools	IS1G21D	4. Every day		
Computer use: Completing [worksheets] or exercises	IS1G21E	Missing	Mode,1	
Computer use: Organizing your time and work	IS1G21F			
Computer use: Writing about your learning	IS1G21G			
Computer use: Completing tests	IS1G21H			
Computer use: [Language arts: test language]	IS1G22A	0. Never		
Computer use: [Language arts: foreign and other national languages]	IS1G22B	1. Less than once a month 2. At least once a month but not every week	Copy,0	
Computer use: Mathematics	IS1G22C	3. At least once a week but not every day		
Computer use: Sciences (general science and/or physics, chemistry, biology, geology, earth sciences)	IS1G22D	4. Every day Missing	Mode,1	
Computer use: Human sciences/Humanities (history, geography, civics, law, economics, etc.)	IS1G22E			PCA
Computer use: Creative arts (visual arts, music, dance, drama, etc.)	IS1G22F			
Computer use: [Information technology, computer studies, or similar]	IS1G22G			
Computer use: Other (practical or vocational subjects, moral/ethics, physical education, home economics, personal and social development)	IS1G22H			
School learning: Providing references to internet sources	IS1G23A	1. Yes	00	
School learning: Accessing information with a computer	IS1G23B	2. No	10	
School learning: Presenting information for a given audience or purpose with a computer	IS1G23C	Missing	01	
School learning: Working out whether to trust information from the internet	IS1G23D			PCA
School learning: Deciding what information is relevant to include in school work	IS1G23E			
School learning: Organizing information obtained from internet sources	IS1G23F			
School learning: Deciding where to look for information about an unfamiliar topic	IS1G23G			
School learning: Looking for different types of digital information on a topic	IS1G23H			

Table D.2: Student background variables used for conditioning (contd.)

Variable	Name	Values	Coding	Regressor
Teaching: Communicating over the internet	IS1G24A	1. I mainly taught myself	00000	
Teaching: Creating documents for school work	IS1G24B	2. My teachers	10000	
Teaching: Changing computer settings	IS1G24C	3. My family	01000	PCA
Teaching: Finding information on the internet	IS1G24D	4. My friends	00100	
Teaching: Working in a computer network	IS1G24E	5. I have never learned this	00010	
		Missing	00001	
Do well: Search for and find a file on your computer	IS1G25A	1. I know how to do this	000	
Do well: Use software to find and get rid of viruses	IS1G25B	2. I could work out how to do this	100	
Do well: Edit digital photographs or other graphic images	IS1G25C	3. I do not know how to do this	010	
Do well: Create a database (for example, using [Microsoft Access ®])	IS1G25D	Missing	001	
Do well: Create or edit documents (for example, assignments for Do well: school)	IS1G25E			
Do well: Search for and find information you need on the internet	IS1G25F			PCA
Do well: Build or edit a webpage	IS1G25G			
Do well: Change the settings on your computer to improve the way it operates or to fix problems	IS1G25H			
Do well: Use a spreadsheet to do calculations, store data, or plot a graph	IS1G25I			
Do well: Create a computer program or macro (for example, in [Basic, Visual Basic])	IS1G25J			
Do well: Set up a computer network	IS1G25K			
Do well: Create a multimedia presentation (with sound, pictures, or video)	IS1G25L			
Do well: Upload text, images, or video to an online profile	IS1G25M			

Table D.2: Student background variables used for conditioning (contd.)

Variable	Name	Values	Coding	Regressor
Computer experience: It is very important to me to work with a computer.	IS1G26A	1. Strongly agree 2. Agree	0000 1000	
Computer experience: Learning how to use a new computer program is very easy for me.	IS1G26B	3. Disagree 4. Strongly disagree	0100 0010	
Computer experience: I think using a computer is fun.	IS1G26C	Missing	0001	
Computer experience: I have always been good at working with computers.	IS1G26D			
Computer experience: It is more fun to do my work using a computer than without a computer.	IS1G26E			
Computer experience: I use a computer because I am very interested in the technology.	IS1G26F			PCA
Computer experience: I know more about computers than most people of my age.	IS1G26G			
Computer experience: I like learning how to do new things using a computer.	IS1G26H			
Computer experience: I am able to give advice to others when they have problems with computers.	IS1G26I			
Computer experience: I often look for new ways to do things using a computer.	IS1G26J			
Computer experience: I enjoy using the internet to find out information.	IS1G26K			

APPENDIX E:

Transformation parameters for ICILS questionnaire scale

Table E.1: Transformation parameters for ICILS questionnaire scale (means and standard deviations of original IRT scores)

Scale name	Mean	SD
Student questionnaire scales		
S_USEAPP	-0.96	1.13
S_USEINF	-0.80	0.90
S_USECOM	0.21	1.04
S_USEREC	0.37	1.14
S_USESTD	-0.72	1.37
S_USELRN	-1.49	1.75
S_TSKLRN	1.27	1.64
S_BASEFF	2.31	1.39
S_ADVEFF	0.52	1.46
S_INTRST	1.77	1.54
Teacher questionnaire scales		
T_EFF	2.40	1.73
T_USEAPP	-2.74	2.16
T_USELRN	-1.82	2.55
T_USETCH	-1.54	2.60
T_EMPH	-0.67	3.31
T_VWPOS	1.65	1.96
T_VWNEG	0.07	1.57
T_RESRC	-0.43	1.93
T_COLICT	0.60	2.10
School questionnaire scales		
C_ICTRES	0.97	1.52
C_HINHW	-0.13	1.54
C_HINOTH	0.38	1.56
P_VWICT	3.34	1.95
P_EXPLRN	1.34	1.98
P_PRIORH	2.16	1.71
P_PRIORS	1.88	1.67

The IEA International Computer and Information Literacy Study (ICILS) 2013 studied the extent to which almost 60,000 lower-secondary students in more than 3,300 schools in 21 education systems worldwide had developed the computer and information literacy (CIL) they need to participate effectively in the digital age. The study investigated differences within and across the participating countries with respect to provision of CIL-related education and students' CIL outcomes. It also looked systematically at associations between those outcomes and student characteristics (e.g., familiarity with using computers, self-reported proficiency in using computers, and home and personal backgrounds) as well as aspects of schools, education systems, and teaching.

The data-collection instruments consisted of a computer-based student assessment (test) and several questionnaires—student, teacher, principal, and school ICT coordinator. A fifth questionnaire, the national contexts survey, was used to gather contextual data from the ICILS national research center in each country.

ICILS 2013 was conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) and builds on a series of earlier IEA studies focusing on ICT in education. ICILS 2013 is one of the more than 30 comparative research studies the association has conducted during the past 50 years. These studies focus on educational policies, practices, and student outcomes in various school subjects taught in about 100 countries around the world.

This technical report for IEA ICILS 2013 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 to national languages and translation verification, as well as 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.

Researchers in the field can use the IEA ICILS technical report to evaluate the content of the published reports, monographs, and articles based on the ICILS data. They can also use it when conducting their own secondary analyses of the data included in the ICILS international database, which is available, along with the database user guide, from IEA.

ACER
The Australian Council *for* Educational Research

DPC
IEA DATA PROCESSING AND RESEARCH CENTER

IEA Secretariat