unesco

Institute for Statistics
Technical Cooperation
Group

4 QUALITY EDUCATION

January 2022

Guidelines for Data Collection to Measure SDG 4.4.2

Institute for Statistics

GUIDELINES FOR DATA COLLECTION TO MEASURE SDG 4.4.2[1]

[1] This paper has been prepared by Andrés Sandoval-Hernández (University of Bath), Maria Magdalena Isac (KU Leuven), Diego Carrasco (Pontificia Universidad Católica de Chile), Daniel Miranda (Pontificia Universidad Católica de Chile) as part of the UNESCO Institute for Statistics (UIS) methodological agenda.

Contents

List of tables

List of figures

Introduction

When the UN Member States adopted the 2030 Agenda and its 17 Sustainable Development Goals (SDGs), there was not much discussion about how these goals were going to be measured. With just under ten years left to achieve the SDGs, world leaders at the SDG Summit in September 2019 called for a Decade of Action and delivery for sustainable development. The Decade of Action calls for accelerating sustainable solutions to all the world's biggest challenges —ranging from poverty and gender to climate change, inequality and improving the quality of education for all. So, deciding on and implementing a measurement strategy for all SDGs and their targets has become a pressing issue.

In this document, we provide guidelines to apply a recently developed strategy for assessing thematic indicator 4.4.2:

- *Indicator 4.4.2 Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills*

This measurement strategy builds on the data provided by International Large-Scale Assessments (ILSA) (Sandoval-Hernández, Isac, & Miranda, 2019; Sandoval-Hernández & Carrasco, 2020; Sandoval-Hernandez, Osorio-Saez & Eryilmaz, 2021). ILSAs are a natural fit for assessing this particular thematic indicator because existing studies have already collected much of the relevant information. Studies like the OECD Programme for the International Assessment of Adult Competencies (PIAAC) are well suited for providing a proxy measurement of Indicator 4.4.2. PIAAC provides high coverage for the concepts considered in this indicator, incorporates them naturally in its frameworks, collects comparable data consistently (allowing long-term monitoring), and has unrivalled data quality assurance mechanisms in place (ensuring data accuracy, validity and comparability).

The strategy has already been applied to the three cycles of PIAAC and allowed to produce scores to measure Indicator 4.4.2 for 40 countries. The scores are available on the UIS database. While having data to measure and monitor these indicators in 40 countries is a significant achievement, it is important to acknowledge that more than two-thirds of UN members do not participate in PIAAC.

For this reason, we have prepared this document, which objective is to offer robust and easy-to-use guidelines, containing detailed technical guidelines for countries that have not participated in PIAAC to collect the data necessary to produce the information that will allow them to measure and monitor SDG Indicators 4.4.2. More importantly, by following these guidelines countries will be able to produce information that is comparable with that of the 40 countries for which this data already exists.

These guidelines heavily rely on two previous reports in which we propose (Sandoval-Hernández et al., 2019) and implement (Sandoval-Hernández & Carrasco, 2020) a measurement strategy for Indicators 4.7.4 and 4.7.5; as well as on the measurement strategy proposal developed ah-hoc for Indicator 4.4.2 (Sandoval-Hernandez et al., 2021) and on the guidelines to collect data for Indicators 4.7.4 and 4.7.5 (Sandoval-Hernandez, Isac, Carrasco & Miranda, 2021). It also incorporates a number of materials that have been produced by different authors and organisations to introduce key concepts in the area of student assessment, review the evidence on their effectiveness, and provide practical insights to produce national assessments of educational achievement (e.g. Anderson & Morgan, 2008a; Greaney & Kellaghan, 2008, 2012; Kellaghan et al., 2009; Lietz et al., 2017; Rutkowski et al., 2014; Shiel & Cartwright, 2015). We also include relevant information from the technical manuals and user guides of TIMSS (Martin et al., 2020), PISA (OECD, 2021), ICCS (Schulz et al., 2018a) and PIAAC (OECD, 2012), particularly the instruments or background questionnaires and their sampling strategy. When one of the chapters is mainly based on one or several of these documents, we indicate it, so the reader can consult those materials to obtain further details.

Apart from this introduction, these guidelines are organised around ten chapters. In the first one, we define what a national assessment is, its main elements and discuss a list of the questions that the assessment described in these guidelines can answer. In the second, we present the decisions that have to be made in preparation for conducting a national assessment. In the third, we introduce the assessment framework used by the measurement strategy for Indicator 4.4.2, and how this framework maps into the instruments of PIAAC. Chapters four and five contain the procedures to be followed to produce a manual for the administration of the assessment, choosing the administrators and ensuring the quality of the data collected. The next chapter provides instructions for selecting a nationally representative sample of schools and students. Chapter seven focuses on the logistics of the assessment and chapter eight on the preparation, validation and management of the data collected. Finally, the last two chapters introduce the procedures to produce the scores and present the results of the measurement strategy.

1. National and International Assessments

National assessments are designed to describe the achievement of students in a curriculum area aggregated to provide an estimate of the achievement level in the education system as a whole at a particular age or grade level (Greaney & Kellaghan, 2008). International large-scale assessments (ILSAs) share the same objective, but their main characteristic is that the assessment is standardized to be conducted in more than one country, in a way that their results can be validly compared. Normally, these assessments involve the administration of achievement tests to a sample of students, usually focusing on a particular sector in the system (e.g. 8th grade in TIMSS and ICCS or 15-year-old students in PISA). Teachers and others (for example, parents, principals, and students) are normally asked to provide background information, usually in questionnaires. When related to student achievement, this background information can provide insights about how achievement is related to factors such as family socioeconomic status, levels of teacher training, teachers' attitudes toward curriculum areas, teacher knowledge, and availability of teaching and learning materials. Note that the guidelines provided in this document will focus not on the student achievement test but on the background questionnaires. More information about this point can be found in Chapter 3, where the assessment framework and the instruments of this measurement strategy are introduced.

To provide statistically valid results, in sample based-assessments like PIAAC, a representative sample of individuals (usually around 5,000) is drawn from each country. More details about the sampling strategy in (inter)national assessments like PIAAC can be found in Chapter 6. Although the best-known ILSAs feature a number of similarities, there are also some substantial differences that need to be considered when comparing the results for different countries (see Rocher & Hastedt, 2020 for a detailed discussion on this point).

What information is produced by international assessments?

Coming back to the similarities among assessments, according to Kellaghan and Greaney (2001, 2004), international assessments seek answers to one or more of the following questions:

- How well are individuals learning specific skills in the country (with reference to general expectations, aims of the curriculum, preparation for further learning, or preparation for life)?
- Does evidence indicate particular strengths and weaknesses in individuals' knowledge and skills?
- Do particular subgroups in the population perform poorly? Do disparities exist, for example, between the achievements of (a) males and females, (b) individuals in

urban and rural locations, (c) individuals from different ethnic groups, or (d) individuals in different regions of the country?

- What factors are associated with achievement? To what extent does achievement vary with characteristics of the learning environment (for example, school resources, teacher preparation and competence, and opportunitied for life-long learning) or with indoviduals' home and community circumstances?
- Are government standards being met in the provision of resources (for example, textbooks, teacher qualifications, and other quality inputs)?
- Do the achievements of individuals change over time?

The guidelines contained in this document will produce information to address most of these questions. The assessment described here can produce information about the proportion of individuals in a given population who reach the targets suggested not by a curriculum but by SDG 4.4.2. The scales or scores used to measure Indicator 4.4.2 can also be estimated for subgroups of the population (i.e. males/females, urban/rural, high/low SES); so, information about disparities can also be obtained. Due to the systematic application of the ILSAs, there is also the possibility to have information to compare with other countries at different time points; and of course, the assessment that we describe here can also be applied to the same cohort at different time points. This question may be of particular interest if new policies are being implemented. It is important, however, to note that these guidelines refer to the application of PIAAC achievement tests and that the items used in the test are kept confidential. So, access to the test items would need to be negotiated with the OECD.

What are the main phases of an international assessment

For international assessments to produce high-quality information, they need to be of high quality, technically sound, have a comprehensive communication strategy and be useful for education policy. To achieve this aim, different authors and organisations consider different key phases that need to be considered in the implementation of high-quality educational assessments. Lietz and colleagues (2017), for example, consider that there are 13 key phases, while Greaney and Kellaghan (2008) consider 16 and the IEA organizes its studies in 10 main steps (2017). All these categorisations include the same key phases and differ only in the way these are organized. Figure 1 shows a synthesis of them and the chapters of these guidelines where each phase is discussed.

Figure 1. Phases of an educational assessment


Intro + Ch. 1
Policy goals
Intro + Ch. 1
Assessment framework
Chapter 3
Instrument development
Chapter 3
Translation verification
Chapters 4 & 5
Technical standards & operations
Chapter 5
Sample design
Chapter 7
Data collection
Chapter 8
Data management
Chapter 9
Scaling methodology
Chapter 10
Reporting and dissemination

2. Decisions to be made for the national assessment

Who should carry out the assessment?

In each country, the ministry of education should preferably endorse the assessment by expressing an interest in monitoring the learning outcomes to be achieved under SDG thematic indicator 4.4.2 and by giving an endorsement to the current measurement strategy.2 The organization in charge may appoint a national steering committee (NSC) to oversee the work and ensure that the achieved results can play a role in future policymaking.

The composition of the NSC is at the discretion of the organisation in charge and may vary from country to country depending on the power structure within the educational system. The NSC may include representatives of different ministries as well as other stakeholders identified as target groups for the dissemination and use of results such as teacher trainers, school inspectors, curriculum personnel, representatives of international and national NGO's etc. The NSC will provide overall guidance and oversee the work of an implementing agency (IA) that will be appointed by the ministry (when necessary in consultation with other structures such as provincial authorities) to carry out the assessment.

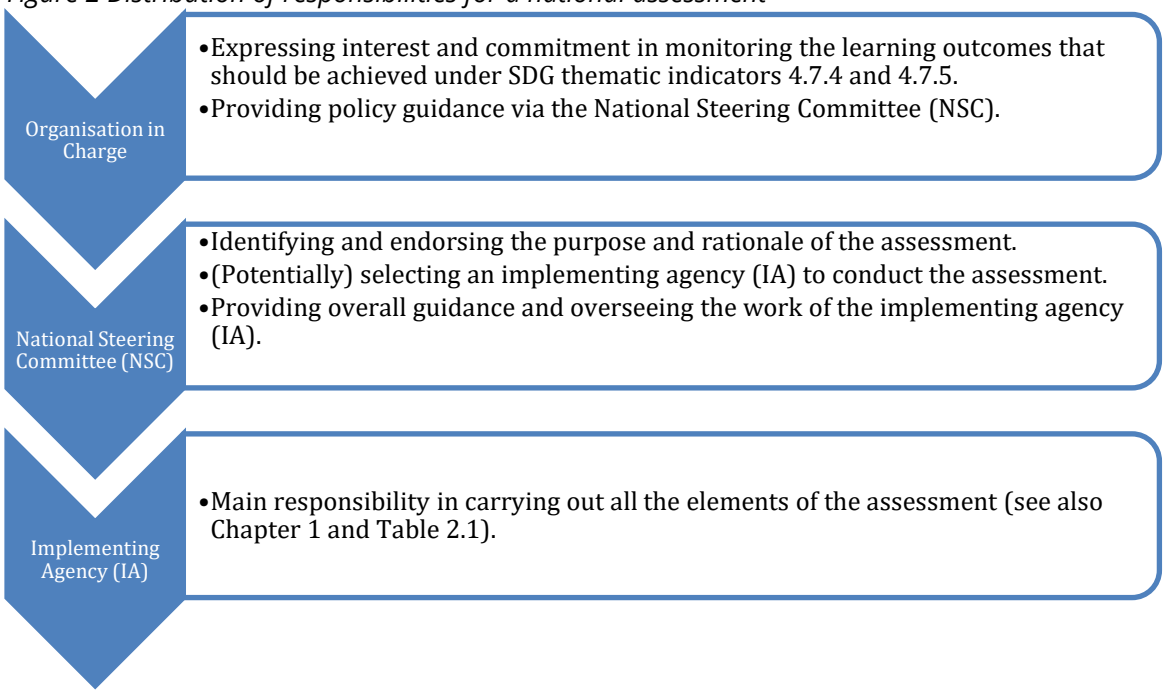
The IA should be a team with proven technical expertise and credibility in organizing large-scale assessments. Various countries organizing national and international assessments often assign this responsibility to different types of groups. These can be for example: a) a team set up within the ministry of education or a public examination agency supported by the ministry of education, b) an autonomous research team working in a university or research center, c) an autonomous international organization with experience in large-scale educational assessment (e.g. IEA, OECD), or d) a team set up within the ministry or an autonomous research team working in a university or research centre, which receives the support of an autonomous international organization with experience in large-scale educational assessment. The decision often involves a reflection on several aspects such as the technical capacity of the IA, the credibility of the IA for

[2] See also: Sandoval-Hernandez, A, Osorio-Saez & Eryilmaz (2021), Sandoval-Hernandez, A., Isac, M.M. & Miranda, D. (2021); Sandoval-Hernandez, A. & Carrasco, D. (2020); and the UNESCO Institute for Statistics official data repository: http://data.uis.unesco.org/

different stakeholders, the costs components associated with each choice, and other administrative and political circumstances3.

The IA will have the main responsibility in carrying out the assessment preferably under the guidance of the ministry of education via the NSC (see Figure 2). Given that the IA will have the main role in carrying out the assessment, the level of technical capacity should be the main criterion in deciding who should be given this responsibility. Table 1 presents a non-exhaustive list of the potential tasks and skills sets that are required to carry out the assessment and should be considered when judging the technical expertise of an IA.

Figure 2 Distribution of responsibilities for a national assessment



[3] For a detailed analysis of advantages and disadvantages of different categories of implementation agencies, please refer to Greaney, V., & Kellaghan, T. (2008, p. 26). Available from: https://openknowledge.worldbank.org/handle/10986/6904

Table 1 Implementing agency (IA): potential tasks and skills sets required

Potential tasks	**Required skills and experience**
• Organizing staff, coordinating and scheduling activities, interacting with different stakeholders (e.g. policy makers, schools and teachers) • Translating and adapting the assessment framework and questionnaires. • Developing manuals for questionnaire administration. • Providing training to test administrators. • Creating a sampling frame. • Contacting and coordinating work with schools. • Collecting data. • Data entry, data management and cleaning. • Statistical analyses (e.g., computing survey weights, producing estimates). • Drafting and disseminating results for different audiences.	• Strong managerial, financial and communication skills (especially from team leader). • High knowledge of the theoretical framework guiding the assessment. • Good organization skills. • High implementation and operational skills. • Experience in working with schools and young people. • Collaboration skills. • Advanced statistical and analytical competence in selecting samples, computing survey weights, preparing data files, producing estimates etc. • Flexibility, openness to learning new methodological approaches. • Ability to communicate findings to different audiences.

Source: Own elaboration, partially based on Greaney & Kellaghan (2008), p. 28-29.

What Population Will Be Assessed and How Frequently?

The population to be assessed

In all national and international assessments, the population to be assessed should be determined by the aims of the assessment and the corresponding information needs. In this assessment, the aim is to collect the data necessary to produce the information that will allow each country to measure and monitor SDG Indicator 4.4.2 and compare this information with the outcomes of the 40 countries for which data already exists (see Chapter 1).

The population to be assessed is therefore defined by the current operationalization of the SDG 4.4.2 as endorsed by UNESCO's Technical Cooperation Group on the Indicators for SDG 4-Education 2030 (TCG) and published in the UNESCO Institute for Statistics official data repository (see: SDG / Goals 1 and 4 / SDG4 / Indicator 4.4.2): http://data.uis.unesco.org/):

- *4.4.2 Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills*

The content of SDG 4.4.2 defines the population to be assessed as "youth/adults". In PIAAC the target population consists of all noninstitutionalized adults between age 16 and 65 (inclusive) who reside in the country (usual place of residency is in the country) at the time of data collection. Adults were to be included regardless of citizenship, nationality or language (see Mohadjer et al., 2013 for more details). Moreover, the operationalization of the indicators implies that the data to be collected should be used to provide information about the overall performance of the target population and not to provide results for each individual in the population (see Chapter 6 for further details).

Given the aims of the assessment, their operationalization and the definition of the target population, it is not necessary to obtain data for each individual in the population (e.g., census-based approaches). The inferences of interest can be obtained instead from a suitably designed high-quality sample (a sample-based approach; see also Chapter 1 and 6). The sample-based approach has a series of advantages. Factors that favour a sample-based approach include: substantially reduced costs in test and questionnaire administration, greater accuracy due to the increased possibility to monitor the quality of implementation, and less time for cleaning and managing data as well as for data analysis and reporting (Greaney & Kellaghan, 2008). Nevertheless, while a sample-based approach provides the means to carry out assessments in an affordable manner, considerable attention to detail is required in designing and selecting the samples.

In this document (see Chapter 6) we refer to a detailed example of a sample-based

approach applied in international large-scale assessments. We particularly elaborate upon the sampling procedure used by OECD's PIAAC. PIAAC provided the sources of data and information that was used to produce the scores of the countries for which data is already available4 (see also Chapter 9). If countries want to produce information that is comparable with that of the 40 countries for which this data already exist, it is advisable that they follow the same procedures as implemented in PIAAC. The reader is referred to Chapter 6 for an in-depth overview of the aspects that are crucial to reflect and decide upon when implementing the recommended sample-based approach including: a precise definition of the target population, an assessment of the population coverage, sample size requirements and sample design etc.

The frequency of the assessment

The frequency of international assessments tends to vary from study to study. The Programme for International Student Assessment (PISA), for example, is implemented every three years; the IEA Trends in International Mathematics and Science Study (TIMSS) every four years; and IEA International Civic and Citizenship Study (ICCS) uses seven-year cycles. The frequency of the assessment should also be determined by its aims. When the purpose of the assessment is to provide information on the performance of the target population on certain indicators (here defined by the content of SDG 4.4.2), one should take into account that this kind of information does not change rapidly. Excessively frequent assessments may fail to register any change and prove to be an unnecessary cost (see also Greaney & Kellaghan, 2008).

What Are the Cost Components of an Assessment?

The cost of an assessment will vary greatly from one country to another depending on the salary levels of personnel and the cost of different services (Greaney & Kellaghan, 2008). A realistic budget is nevertheless essential for the success of the assessment. At the beginning of the project, the different stakeholders should assess the budget needs in consultation with assessment experts and financial decision-makers from the ministry and/or the implementing agency.

Although no established formula exists, it can be useful to have an overview of the potential cost components based on the various phases of the project, the actors and the tasks involved. A non-exhaustive list tailored to the assessment proposed in this document may include the following components:

[4] UNESCO Institute for Statistics official data repository (see: SDG / Goals 1 and 4 / SDG4 / Target 4.4.2): http://data.uis.unesco.org/

National Steering Committee (NSC). Costs related to establishing the NSC and associated activities such as recruiting participants and organizing meetings.

Implementing agency. Costs related to eventual personnel needs and providing facilities and technical equipment.

Designing the assessment framework and instruments/questionnaires. In the current case, this category of costs is greatly reduced due to the fact that an assessment framework is already developed and questionnaires are adapted from existing instruments (see Chapter 3). Nevertheless, budgetary provisions should be made for activities related to translating and adapting this framework and instruments to the specific language and context of each country. Personnel needs (experts), facilities and technical equipment required should be considered.

Sampling procedures. Costs related to expert personnel responsible for creating the sampling frame and drawing the sample of schools and students (see also Chapter 6).

Administration and data collection. Data collection is by far the most expensive component of any assessment. In some countries it may take up to 50 percent of the budget (Greaney & Kellaghan, 2008). It involves many tasks such as recruiting and training questionnaire administrators, designing questionnaire administrators' manuals, designing, administering and retrieving the questionnaires (either in print or online) and ensuring efficient contact with schools (see also Chapters 4, 5 and 7).

Data preparation, validation and management. Costs related to the production of codebooks, data management, verification and cleaning must be handled by expert personnel with access to necessary equipment (see also Chapter 8).

Data analysis and reporting. Costs related to computing and reporting different estimates (e.g., survey weights, indicator scores and thresholds) must be handled by expert personnel with access to necessary equipment (see also Chapter 9).

Reporting and follow-up activities. Costs related to the communication and dissemination of findings to different audiences such as the production of policy briefs or training for different stakeholders in interpreting and acting on the findings (see also Chapter 10).

When considering costs and if possible, countries may also draw information from budgets developed for conducting other international assessments such as PISA, TIMSS or ICCS5 in their country or in countries with comparable conditions in terms of salary levels of personnel and price of different services. Nevertheless, it should be taken into

[5] For information related to the cost of the ICCS study please refer to: https://www.iea.nl/publications/international-study-participation-fees-iccs-2022

account that the scope of the particular assessment proposed in this document is much smaller than the one of any of these surveys. The framework and instruments are already designed, and the content of the questionnaire is significantly shorter compared with the other assessments (see also Chapter 3). Therefore, the costs associated with this proposed assessment meant to measure and monitor SDG Indicator 4.4.2 will most likely be lower.

3. The assessment framework and instruments

Most international assessments are directed at measuring a set of cognitive or non-cognitive outcomes that are important for providing information on the performance of the educational system on certain indicators. In the current proposal, the assessment is designed to assess the performance of countries on SDG Indicator 4.4.2. Similar to other national and international assessments, providing an appropriate *assessment framework* is extremely important. The *assessment framework* clarifies in detail what is being assessed, why it is being assessed, and how it is being assessed. The definition of concepts and their operationalization provides guidance for elaborating/selecting the assessment instruments and analyzing and interpreting the results. The assessment framework usually includes two main components: the purposes and the definition/s guiding the assessment and the operationalization of the main concepts, which is then used to elaborate a measurement strategy, design or select the appropriate assessment instruments and guide the interpretation of the findings.

In this document, we aim to provide guidelines to apply a recently developed measurement strategy for assessing SDG Indicator 4.4.2 using information and guidance from ILSAs in education such as PIAAC. In what follows, we discuss the main components of the assessment framework as elaborated in previous work and for the purpose of this document.

Background: why is it being assessed

In September 2015, UN Members formally adopted the 2030 Agenda for Sustainable Development at the United Nations Sustainable Development Summit in New York. The Sustainable Development Goals (SDGs) are a call for action by all countries to promote prosperity while protecting the planet. They recognize that ending poverty must go hand-in-hand with strategies that build economic growth and address a range of social needs including education, health, social protection, and job opportunities while tackling climate change and environmental protection.

The Agenda contains 17 goals including a global education goal (SDG4). SDG4 establishes that by 2030 we have to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" and has seven targets and three means of implementation. One of these targets, 4.7, refers to the knowledge and skills that are necessary for a sustainable future.

Target 4.7 By 2030, ensure that all learners acquire the knowledge and skills needed to promote sustainable development, including, among others, through

education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development.

Among others, target 4.7 includes the following thematic indicator:

> **4.4.2** Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills.

In this document, we aim to describe and implement a measurement strategy for this thematic indicator using data from OECD PIAAC. To do so, we build on a report[6] previously published by the Global Alliance to Monitor Learning (GAML) describing a proposal of a measurement strategy for this indicator (see also Sandoval-Hernández et al., 2019). This report establishes a global content framework for indicator 4.4.2 and carry out a mapping exercise to evaluate the extent to which the different concepts contained in the framework can be operationalized with the instruments and procedures of OECD PIAAC.

Definition of concepts: what is being assessed?

To arrive at our operational definitions we build on the Digital Literacy Global Framework (Law et al., 2018) and the Recommendations on Assessment Tools for Monitoring Digital Literacy (Laanpere, 2019). Drawing on this body of literature we use the following working definition of Digital Literacy (DL):

> ***Digital Literacy (DL)***
> Digital literacy involves the confident and critical use of a full range of digital technologies for information, communication and basic problem-solving in all aspects of life. It is underpinned by basic skills in information and communication technology (ICT): the use of computers to retrieve, assess, store, produce, present and exchange information, and to communicate and participate in collaborative networks via the Internet.

[6] Measurement Strategy for SDG Global Indicator 4.4.2 Using International LargeScale Assessments (Sandoval-Hernandez et al., 2021).

Operationalization of concepts: what precisely is being assessed?

Based on the two reports mentioned above, we establish a Global Content Framework for indicator 4.4.2. This exercise resulted in a framework with seven competence areas and several competences within each area (see Table 1). The main competence areas are Devices and software operations, Information and data literacy, communication and collaboration, Digital content creation, Safety, Problem-solving, and Career-related competences.

Table 2. Global Content Framework for SDG indicators 4.4.2

Competence areas	Competences
0. Devices and software operations	0.1 Physical operations of digital devices
	0.2 Software operations in digital devices
1. Information and data literacy	1.1 Browsing, searching and filtering data, information and digital content
	1.2 Evaluating data, information and digital content
	1.3 Managing data, information and digital content
2. Communication and collaboration	2.1 Interacting through digital technologies
	2.2 Sharing through digital technologies
	2.3 Engaging in citizenship through digital technologies
	2.4 Collaborating through digital technologies
	2.5 Netiquette
	2.6 Managing digital identity
3. Digital content creation	3.1 Developing digital content
	3.2 Integrating and re-elaborating digital content
	3.3 Copyright and licences
	3.4 Programming
4. Safety	4.1 Protecting devices
	4.2 Protecting personal data and privacy
	4.3 Protecting health and well-being
	4.4 Protecting the environment
5. Problem-solving	5.1 Solving technical problems
	5.2 Identifying needs and technological responses
	5.3 Creatively using digital technologies
	5.4 Identifying digital competence gaps
	5.5 Computational thinking
6. Career-related competences	6.1 Operating specialised digital technologies for a particular field
	6.2 Interpreting and manipulating data, information and digital content for a particular field

Source: Sandoval-Hernandez, Osorio-Saez & Eryilmaz (2021)

Assessment instruments: how is it being assessed?

In previous work (Sandoval-Hernandez et al., 2021), we carried out a mapping exercise to evaluate the extent to which the different concepts contained in the framework (i.e., competence areas and competences) can be operationalised with the instruments and procedures of existing digital literacy assessments. The digital literacy assessments evaluated were: OECD's Programme for the International Assessment of Adult Competencies (PIAAC) (OECD, 2012), the OECD's Programme for International Student Assessment (PISA) (OECD, 2019), and the IEA International Computer and Information Literacy Study (ICILS) (Fraillon et al., 2019).

The mapping exercise identified PIAAC as the most valuable source of information for SDG Indicator 4.4.2. PIAAC was chosen due to its conceptual framework (OECD, 2012), which showed the highest coverage of the topics relevant to this indicator. Additional reasons for the selection of PIAAC were that its target population covers the two groups mentioned in the indicator (youth and adults), as well as its potential to inform long-term monitoring.

Appendix I includes sample items from the PIAAC instruments that were used to evaluate Problem-Solving in Technology-Rich Environments (PSTRE), which is the construct that we identified as having large overlaps with SDG 4.4.2 Glocal Content Framework (see Sandoval-Hernandez et al., 2021 for more details).

Based on these instruments and on the available data, a series of measurement models using items from PIAAC can be estimated in order to generate scores (i.e. percentage of individuals meeting the indicator). Moreover, this information can be also used to identify proficiency levels of individuals based on each respective score. For an in-dept overview of the required procedures the reader should consult Chapter 9.

4. Designing a manual for questionnaire administrators

A manual is required to guide the questionnaire or test administration, which must be standardized so that all students participate in the assessment under the same conditions. All the recommendations presented in this chapter are based on four manuals that compile different relevant aspects for this report (Anderson & Morgan, 2008a; Greaney & Kellaghan, 2012; Lietz et al., 2017; Schulz et al., 2018b). In the reminding of this section, we answer some common questions related to the development of a manual for questionnaire administration, including what is a manual, what is it for, and the sections that ir normally should include. We finally list some good practices taken from the experience of different implementation agencies around the world.

What is a Manual for Test Administration?

A manual for questionnaire or test administration is a document that describes the different steps and responsibilities that are needed for an educational assessment under standardized conditions for all students in a given sample. A good manual contains all necessary information and is easy to use. The information is logically ordered, instructions are clear and complete, and language is simple and direct. Bullet points, boxes, or tables will make the information easier to read.

In the interest of efficiency and to limit the number of documents test administrators have to carry, the key information related to timing, student preparation, packing and returning of tests and questionnaires, and instructions for administration should be included in one document—the test administration manual. Instructions that are read aloud to pupils should be in large, bold print. A person entrusted with training test administrators should go through the entire manual with at least a sample of test administrators prior to formal training of the selected administrators. No matter how well they claim to be qualified, test administrators should not be left to go through the manual on their own.

What is this manual for?

The main purpose of the manual is to specify the exact conditions under which a test must be conducted, including preparation requirements and procedures for ensuring test security. Students taking the assessment must work through the same practice questions and receive the same instructions about how to show their answers. All must be given the same amount of time to complete the questionnaire with the same degree of supervision.

Students' performance on a national assessment should be a measure of their ability to answer the items without external support or to collect their opinions, feelings

or beliefs. The students should understand what they have to do and how to show their answers, but they should not be given any other assistance or have access to any resources that are not a part of the assessment. Following the procedures laid down in an administration manual should help ensure that this will be the case.

What sections should be included?

The administration manual should provide information answering each of the following questions:

Table 3. Contents of the administration manual

Key question (sections)	**Examples**
What is the test for?	Brief explanation of the purpose of the test and the way the data will be used
Which tests are given, which students are tested, and when are they tested?	Details about which test, length administration of each, which students, dates and times, required breaks or any flexibility option for the administration.
What test materials are needed?	List of all the test supplied materials that are supplied, quantities per students, per teacher and per school (i.e. pencils, erasers) .
How should the room be set up for the test?	Description of physical facilities needed and description of resources that must be removed/covered (i.e. number of desks, covering up posters with grammatical rules, etc.)
What preparation is required?	Description of motivation for staff members, required information, instructions for booklets organization, organization of students, etc.
How should the test be conducted?	Description of procedures for booklets administration. For instances, registry of information, check procedure, practices questions administration, instructions for students, how much test must take, conditions for administration, rules for people allowed into the room, etc.
How should test materials be stored?	Procedures to ensure the security of the test materials before, during, and after the test.
Who can be contacted for help?	Contact details for people who can assist with problems or provide additional information.

As can be seen, the manual for test administrations must contain several details to ensure the standardization of the data collection procedures. Any additional information about the management and movement of materials in and out from schools could be included, depending of the needs of each administration agency.

Information about the general conditions of questionnaire administration and the preparation of questionnaire materials should be comprehensive but, at the same time, as brief as posible. The next table show an example of it.

Table 4. Contents of a test administration manual (example from the Department of Education, Papua New Guinea)

Administration Manual Instructions	**Information for Teachers and Principals**
In a national assessment, the following information appeared in a large Font (Arial 14 point), taking up the entire opening page of the administration manual: Please read this Administration Handbook before your students do the test. **Students must do this test over TWO DAYS.** • The test is divided into four sessions. Students must do two sessions each day. • Students must have a break between each session. • Do not let students work through the whole test at once. **Administration Rules** • Teachers must supervise all sessions at all times. • Students must NOT take test booklets out of the classroom or work on them after the teacher has left.èStudents must use the pencils with erasers on the end that have been supplied. • Students must not use any classroom materials, such as workbooks, dictionaries, or calculators, when they do the tests. • Students must not be helped with answering the questions. For example, if a student does not understand what to do, explain the practice questions again and tell him/her to try his/her best but do not give any further help. **Test Security** • The test materials must be STORED SECURELY AT ALL TIMES. • Student test booklets must NOT be copied for any purpose. • Students must NOT take test booklets home.	Information about the test materials should be concise and listed in a way that is easy to check. The following extract from a large-scale assessment in Papua New Guinea tells the head teacher or principal what materials have been sent to the school and how to fi nd out which classes will participate in the test: **Test Materials** Your Senior Primary School Inspector will tell you which classes in your school need to participate in this test. You should have received the following materials: • a cover letter for the head teacher • a student test booklet for each participating student • an administration handbook for each teacher administering the test • a teacher background questionnaire for each participating teacher • a pencil with an eraser on the end for each participating student If any materials are missing or you do not have enough materials, please contact your Senior Primary School Inspector.

Source: Papua New Guinea Department of Education 2004.

Good practices

The manual should be used by the principal of schools (or headteacher) and the test administrator. On the one hand, the principal (or headteacher) needs the manual to ensure his or her school is appropriately prepared for the test administration. On the other hand, test administrators need the manual to tell them exactly what they have to do to administer the test properly and when and how to do it.

The Principal should know enough about the test to encourage the staff and the students to support the administration and to motivate students to try their best. The headteacher (or principal) should have sufficient information to be able to organize the school and to make sure that the correct students are available at the required time, with the right materials; that they will have adequate space to take the test; and that test materials can be stored securely. The test administrator needs to check that sufficient test materials are available and that the correct students have been selected to take the test. They need to know what information to give students about the test, how to explain the practice questions, and how much time students have to do the tests. They also should know what security procedures to use for storing test materials.

There are some good practices recommended to ensure the usability of the manual:

- The manual should be prepared for tryout in the pretest or field test of the test items. Pretesting the manual will highlight any misunderstandings or ambiguities that require clarification or refinement in the final version. Because the pretest or field-test conditions should be as similar as possible to those of the final administration, the manual should be in as finished a form as possible at the time of the tryout.
- General instructions about the administration of the test can usually be written any time after the blueprints have been finalized. The blueprints should specify all the requirements about the number of tests and their length and about which students should take the test.
- During the pretest, the administrator should collect information such as the following to assist the test development manager in refining the final test:
    - Whether students needed all the practice questions, whether there were enough practice questions, and whether explanations were sufficiently clear.
    - Whether the test was the right length or too long, and approximately how many students finished more than 10 minutes early (if different forms are used in the same class, the administrator can compare the length of time students required for each form)
    - Whether students appeared to be engaged by the test.
- The manual should be proofread to ensure instructions for test administration, practices, and conditions for the application are clear for all.

5. The questionnaire administrator

This section characterizes or defines the questionnaire or test administration process, including the selection of administrators, their instructions, quality assurance and a proposed checklist for ensuring the successful completion of the process. The contents of this chapter are mainly adapted from Anderson and Morgan (2008b).

Selection of test administrators

People should be confident that the test was administered under standardized conditions. Test administrators must be widely regarded as trustworthy. The choice of test administrator depends on conditions in a country. In some countries, classroom teachers administer national assessment tests to their own students. More often than not, however, teachers other than those who teach the students who are taking the test or individuals who are external to the school are entrusted with this task. In some countries, data collection is contracted to a body that specializes in that activity. School inspectors may be ideal administrators in some countries but problematic in others. If the inspectors see test administration as an additional task that is outside their job description, that uses scarce resources, or that is of little interest to them, they may not be motivated to do the job properly. External administrators are used in some national assessments. Ideally, they are people who can follow instructions precisely, have the time and resources to do the task properly, and have no particular interest in the outcome of the test other than to administer it correctly. Some possible advantages and disadvantages of using personnel from different backgrounds are summarized in Table 5. It is, however, important to mention that providing clear guidelines and intensive training can help address any disadvantages that may exist.

Table 5. Advantages and disadvantages of using different actors as questionnaire administrators

Category	Advantage	Disadvantage
Teachers	Are professionally qualified	May have difficulty unlearning usual practices (for example, helping students) and learning new ways of dealing with pupils
	Are familiar with the children	May feel they are also being assessed and may try to help the children (if their own class is being assessed)
	May be less expensive than others, especially in terms of travel and subsistence	May be difficult and costly to organize and train
	Are likely to be fluent in the area or local language	
Inspectors and teacher trainers	Are likely to have classroom experience	Might be overly authoritarian
	Will become involved as partners in the national assessment, which may give them an interest in the outcomes	Might be tempted to conduct inspection activities in addition to administering tests
	Are likely to know the location of most schools	Are likely to be more costly than teachers
		May feel they need not follow the detailed instructions in the manual
University students	Are readily available, especially during university vacations	May not be very reliable
	Are likely to follow instructions	May lack the authority required to deal with managers, principals, and others
	Are more likely than others to withstand harsh travel conditions	Are difficult to hold accountable
	Can often use a work opportunity	May not be fluent in the local language
	Are relatively inexpensive	May not communicate a sense of respect and authority in front of students
Assessment or examination board personnel	Are professionally qualified	May be too authoritarian, especially if they are used to supervising public exams
	Are directly accountable to the appointing authority	May lack recent classroom experience and therefore not exude a sense of authority in front of students
	Tend to be reliable	May lack experience at the particular educational level being tested
	Are good at recordkeeping	Are expensive to maintain in the field
	Tend to consult before making major decisions	May not be fluent in the local language

Because faulty test administration tends to be the most common source of error in a national assessment, particular attention should be paid to selecting, training, and supervising test and questionnaire administrators. Above all, persons assigned this position should be trustworthy, responsible, and committed.

Instructions

The manual should distinguish between specific instructions that must be followed to the word from more general instructions that allow the administrator some scope to adapt them to the conditions in the class. Some relevant aspects for instructions are:

- The test administrator should not deviate from any specific instructions. Pretesting the manual should help identify any errors or ambiguities in the instructions.
- Test administrators should help students only to understand what they have to do and how to show their answers.
- If a student asks for help, the administrator should tell the student just to try his or her best. Test administrators should make clear that they cannot help any students answer questions.
- In some tests, administrators may read the questions to students. The test administrator should read slowly and distinctly the whole test aloud to the class, question by question, or read single questions as requested by the students.
- Administrators should ensure that students are aware of the time they have to do a test.. Administrators must have a watch or clock.
- Administrators should quietly encourage students to attempt the whole test.
- Only materials that are specified in the manual are allowed in the room during test administration.
- The test administrator, students participating in the test, and possibly a supervisor should be the only people in the room during test administration. The head teacher or principal or other teachers should not be permitted to walk around the room. The test manager should be notified of unavoidable changes in test administration conditions.
- During the administration of the test, the administrator should collect information about any variations that occur in the conditions of administration for individual students.
- The national assessment team should ensure that each test administrator has, or has access to, a timing device to be used during test administration. The test administrator is responsible for ensuring that teachers do not help students and that students do not copy from each other or bring unauthorized

materials into the room. School conditions will dictate seating arrangement options.

- The test administrator should check that desks are free of books and other materials prior to testing. National assessments that use more than one form of a test reduce the possibility of copying by requiring students seated near each other to take different versions of the test.

The test administrator should complete a student tracking form (See Figure 3 for an example), which is sent to schools with test booklets and questionnaires. Information from this form will be needed at the data cleaning and analysis stages (for example, in weighting data). Information recorded on the tracking form usually includes each student's name, assigned identifier (ID) number, date of birth, gender, and record of attendance at individual testing sessions and, where applicable, replacement sessions. If the testing requires more than one session, the student's presence should be noted for each session.

Figure 3. Example of Student Tracking Form

School name: _____

School ID	Class ID	Class name	Grade

Student name	Student ID	Date of Birth	Excluded	Session			Replacement session		

Source: Anderson and Morgan (2008a)

- The test administrator must ensure that all tests and questionnaires, used and unused, are kept secure and are returned to the national assessment center. This step is important because items, and in some instances an entire test,

might be used in a subsequent national assessment. If some teachers and students have prior access to those items, the credibility of the subsequent assessment would be undermined. The paper or rough notes used by students while doing the tests should also be returned to the national assessment office.

Quality procedures

For consistent administration of testing process for all students' administrators should be selected for their suitability for the task. Next are listed some of criteria for ensuring quality for test administrator:

- They should be fluent in the language in which the manual is written.
- They also should be committed to doing their task well.
- They should attend a training session that explains the purpose of the test and their role in its administration.
- They should understand why following instructions is important, and they should be given the opportunity to practice administering the test with fellow test administrators.
- They should have the opportunity to ask questions about the procedures outlined in the manual.
- If teachers are to administer the tests to their own students, the training must ensure that they understand the purpose of the test and are reassured that the data will not be used to judge them.
- They should understand the importance of not assisting students in answering questions.
- Administrators should be supervised for at least some of the time they administer the test. Supervising everyone may not be possible, but random checks of some administrators should be feasible.
- Administrators can also be asked to fill in and sign checklists of their tasks to help ensure that they have completed their job.

Checklist and common problems

Details of what should be in the administrator's checklist will vary, depending on who is administering the test and the procedures developed for tracking booklets and ensuring security. Table 6 provides an example of an administration checklist used in the Philippines. The idea is that the administrator checks every item to show that he or she completed it and sign the form at the end. A further example can be seen in Greaney and Kellaghan (2012).

Table 6. Administration Checklist: An Example from the Philippines

Name:		Date:	
Task	Reference	Time	Completed
1. Complete the student test booklet allocation (STBA) form by inserting the test numbers in consecutive order and entering the students' names in alphabetical order.	STBA form	10 min.	
2. Administer teacher questionnaire.	Teacher questionnaire form	15 min.	
3. Complete feedback form	Teacher feedback form	10 min.	
4. Distribute the **allocated test to each student** and mark absent against students not in attendance.	STBA form	10 min.	
5. Read introduction from Guidelines.	Administrator Guidelines, p. 7	5 min.	
6. Ask students to complete **student details** on front cover of test.	Administrator Guidelines, p. 9	5 min.	
7. Check that every student has completed the required student details on front cover.		10 min.	
8. Follow instructions for Session 1	Administrator Guidelines, pp. 11–13	60 min.	
9. For breaks, ask students to leave the room by row and to leave their test on their desks.		15 min.	
10. Follow instructions for Session 2.	Administrator Guidelines, pp. 15–17	60 min.	
11. For breaks, ask students to leave the room by row and to leave their test on their desks.		15 min	
12. Follow instructions for Session 3.	Administrator Guidelines, pp. 19–21	70 min	
13. Collect all test booklets and check off their return using the STBA form.	STBA form	10 min	
14. Account for all tests and make sure every test has been returned.		5 min	
15. Dismiss class.	STBA form	2 min	
16. Sign STBA form.	STBA form	2 min	
17. Collect and pack all test materials in the box provided, including: i. STBA form ii. Teacher questionnaire iii.Teacher feedback form iv. All completed tests v. All unused tests.		10 min.	
18. Securely store materials		10 min	
19. Return materials to your senior district supervisor (SDS) for the Regional Assessment of Mathematics, Science, and English (RAMSE).	SDS RAMSE distribution form	Travel time	
20. Return this completed checklist to your SDS.	RAMSE administrative checklist	2 min.	
Administrator signature _____			

Source: Anderson and Morgan (2008a).

6. Sampling, weighting, and variance estimation

The objective of many national assessment programs is to obtain results at the individual, administrative unit, and national levels. Such assessments are normally used to make decisions about individuals' progress towards specific outcomes or about the state of affairs in specific areas. Many of these assessments are based on information from the whole population of interest. In these circumstances, because every individual participates (i.e. census), there is no sampling needed. Therefore, there are no issues of sample design and selection involved, and no issues related to the need to provide analysis weights. In our case, however, the goals of the study do not include the provision of individual results for all the individuals in the population. Rather, the purpose is to make inferences about the whole population only. This purpose can be extended to providing results for a variety of population subgroups, examining the distribution of the variables measured within and across these subgroups.

Given these goals, it is not necessary to obtain data for each individual in the population. The inferences of interest can be obtained from a suitably designed and executed sample of them (Rust, 2014). This, of course, offers the potential to greatly reduce the cost and burden of this assessment. While sampling methods provide the means to carry out assessments in an affordable manner, considerable attention to detail is required in designing and selecting the samples. Furthermore, additional calculations are needed to produce the sampling weights and the variance estimation procedures (replicated weights) that are needed to produce the final estimates. These three topics are covered in this chapter.

Target population and sampling frame[7]

Countries participating In PIAAC were required to develop their sample design and selection plans according to the standards provided in the PIAAC Technical Standards and Guidelines (TSG) and to submit their plans to the Consortium for approval. The sample design plans included information about sampling frames and their coverage, providing descriptions of the national sample designs that included stages of sampling, probabilities of selection, sampling units and sample sizes. The sample selection plans included detailed information about the processes for sample selection at each stage of sampling. In addition, the countries were required to complete and submit quality control sample selection forms to the Consortium to verify that the sample selection was conducted in an unbiased and randomized way consistent with PIAAC standards.

[7] From this point, the information included in this chapter is taken from PIAAc Technical Manual (Mohadjer et al., 2013).

The target population for PIAAC consists of all noninstitutionalized adults between ages 16 and 65 (inclusive) who reside in the country at the time of data collection (more details about the target population are discussed below). In the rest of this chapter, we provide more detail on the PIAAC target population and the national target populations if expanded beyond the PIAAC standard definition. We also describe the sources of country sampling frames and their coverage of the target population. The TSG allowed each country to choose a sample design and selection approach that is most optimal and cost-effective, as long as the design applies full selection probability methods to select a representative sample from the PIAAC target population. Descriptions of the standard PIAAC and examples of national sample designs and probabilities of selection are also presented, as well as the definition of sampling units and sample selection methods. We finally present the PIAAC target sample sizes, describe the process applied to determine the initial sample sizes and a summary of the sampling quality control procedures used in this study.

Definition of the target population

A clear and precise definition of the target population is necessary to ensure that the population of interest is adequately covered by each participating country and to maintain consistency and comparability across countries. The PIAAC target population consists of all non-institutionalized adults between ages 16 and 65 (inclusive) who reside in the country (usual place of residency is in the country) at the time of data collection. Adults were to be included regardless of citizenship, nationality or language. The target population excludes adults in institutional collective dwelling units (or group quarters) such as prisons, hospitals and nursing homes, as well as adults residing in military barracks and military bases. However, full-time and part-time members of the military who do not reside in military barracks or military bases are included in the target population. Adults in other non-institutional collective dwelling units (or group quarters), such as workers' quarters or halfway homes, are also included in the target population. This includes adults living at school in student group quarters such as a dormitory, fraternity or sorority. Adults who were unable to complete the assessment because of a hearing impairment, blindness/visual impairment or physical disability are considered in scope; however, they were excluded from PIAAC response rate calculations because the assessment does not accommodate such situations.

We suggest that countries following these guidelines keep the definition of the target population stipulated in PIAAC technical documentation. As in the original PIAAC study, countries might want or need to have some exclusions or oversample certain groups in the population. This is of course possible as long as there is transparency in the decisions made in this regard. Details on the exclusions and oversampling options taken by the

countries participating in PIAAC can be found in Mohadier et al. (2013) and an example of how this can be reported is included in Table 7.

Table 7. Countries conducting oversampling – PIAAC Round 2

Country	Group oversampled
Israel	The Arab population and Ultra-orthodox
New Zealand	Persons of Maori and Pacific ethnicities; Persons aged 16-25 years
Singapore	Twenty-year-olds who participated in PISA 20091; Foreign professionals who are Employment Pass holders and working in Singapore for a short term

Source: Mohadier et al. (2013)

Sampling frames and their coverage

The sampling frame is the list from which the sample is selected, so the quality of the sampling frame affects the quality of the sample. In addition, adequate information on the frame must be available to conduct sampling, data collection, weighting, and nonresponse bias analyses. In PIAAC, most countries with multiple stages of selection (see below) specified multiple frames. Those frames should, in principle, be reviewed by the implementing agency to ensure they include sufficiently reliable information for sampling individual units and ultimately locating individuals for the interview and assessment. Another important piece of information is the non-coverage rate which, in the case of PIAAC, could not exceed 5%. Thus the sampling frames for each country were required to include 95% or more of the standard PIAAC target population.

Sampling frames

It is fundamental that sampling frames are up to date and include only one record for each member of the target population. Countries have to examine their sampling frames and eliminate duplicate records when lists were combined to create a sampling frame. Countries should assess the extent of duplication and the proportion of out-of-scope units on the frame and, if necessary, develop a plan to correct these problems. In addition, countries should also evaluate and develop plans to address any non-coverage in the frame that was not addressed in the documentation of country-specific. The methodology used to create these frames should also be reviewed by the implementing agency. Multi-stage sample designs require a sampling frame for each stage of selection (see below). Countries can use national population registries as sampling frames, which contain useful variables for stratification, weighting and nonresponse bias analyses. If the country has a list of residents that is of sufficient quality, no frame of households or

household sampling is necessary. An example of the sampling frames used by some countries in PIAAC is shown in Table 8.

Table 8. Examples of sampling frames for countries with population registry samples – PIAAC Round 1

Country	Sampling frame		
	Stage 1	**Stage 2**	**Stage 3**
Austria	Population registry, 2011	n.a.	n.a.
Denmark	Population registry, 2011	n.a.	n.a.
Estonia	Population registry, 2011	n.a.	n.a.
Finland	Statistics Finland's population database (based on the Central Population Register), 2011	n.a.	n.a.
Flanders (Belgium)	Population registry, 2011	n.a.	n.a.
Germany	German Census Bureau frame of communities, 2011	Local population registries, 2011	n.a.
Italy	National Statistical Institute of Italy frame of municipalities, 2011	Household registries held by municipalities, 2011	Population registries, 2011; combined with field enumeration
Japan	Population registry, 2011	Resident registry, 2011	

n.a. Indicates there is no such stage in the country's sample design.
Source: Mohadier et al. (2013)

Non-coverage of the target population

The non-coverage rate suggested for this study is the same that was adopted for PIAAC, no more than 5% (for all the combined all stages of sampling). Therefore, the sampling frames for each country are required to include 95% or more of the standard PIAAC target population. All exclusions to the core PIAAC target population, whether or not they exceed the threshold, should be reviewed by the implementing agency. Exclusions should be acceptable only if they occur because of operational or resource considerations, such as excluding persons in hard-to-reach areas. For the sake of transparency, each country should identify possible exclusions before sample selection. Adjustments for any non-coverage of the target population in each country should be made through benchmarking during the weighting process. A complete list of exclusions should be presented as part of the reporting. Table 9 shows some examples of how this information was reported in PIAAC. Note the non-coverage rate in the tables accounts for excluded subpopulations

such as undocumented immigrants. Other exclusions that occurred as a natural part of the survey process are not included in the expected non-coverage rate.

Table 9. The portion of target population not covered by PIAAC sampling frames of selected countries – Round 1

Country	Percentage of target population not covered	Group not covered
Austria	0.6 %	Undocumented immigrants
Denmark	<0.1%	Undocumented immigrants
Estonia	2.8%+	Persons without a detailed address; undocumented immigrants (no estimate provided)
Finland	0.2%	Undocumented immigrants
Flanders (Belgium)	1.0%	Undocumented immigrants
Germany	0.5%	Undocumented immigrants
Italy	0.8%+	Adults in noninstitutional group quarters; undocumented immigrants (no estimate provided)
Japan	2.2%	Nonnationals; undocumented immigrants
Netherlands	0.9%	Undocumented immigrants

Source: Mohadier et al. (2013)

Sample design

The sample design suggested for this study is based in that of PIAAC. That is a self-weighting design of persons (or of households, for countries without person registries). A self-weighting design is achieved when each sample person (or household, if sampling dwelling units) has an equal probability of selection. For countries that are geographically large, the typical sample design is a stratified multistage clustered area sample. For countries that are geographically small, the sample design may have less clustering and fewer stages of sampling.

Each country should choose a sample design and selection approach that is most optimal and cost-effective, as long as the sample design applies full selection probability methods. Each country is required to produce a probability-based sample, representative of the target population of the country. Probability-based samples are important because they are essential for two main reasons. First, probability sampling encompasses a set of designs that leads to a variety of unbiased sampling approaches that allow analysts to generalize the results to the target population. Second, measures of precision related to survey estimates (i.e., standard errors, margins of error, confidence intervals) can be

computed under a probability design only. Hence, statistical tests for differences between survey estimates are possible only under a probability-based design.

In what comes next, we present examples of the PIAAC standard probabilities of selection as applied to different country's designs. We also present the sample units selected at each stage of selection, as well as the sample selection methods. Finally, we present the factors contributing to the sample size determination and the sample sizes.

Probabilities of selection based on PIAAC standard design

Each person in the target population must have a non-zero probability of selection resulting from the application of established and professionally recognized principles of scientific sampling. That is, every in-scope person must have a chance of being selected into the sample. The following presents the recommended approach for selecting the ultimate sampling unit. We present the cases for one-, two-, and three-stage sample designs, respectively. The approach is based on PIAAC standards and guidelines.

One-stage sample designs

For a one-stage sample design without any explicit stratification, let

n = total number of persons to be sampled, and

N = total number of eligible persons.

The probability of selecting person l is $rr = n/N$.

For a one-stage stratified sample design, let

n_h = number of persons to be sampled in stratum h; and N

N_h = number of eligible persons in stratum h.

Further, let $rr = nn/NN$, then the probability of selecting person ll in strata h is $P_{hl} = rr$.

The sample size is allocated to strata as n

$n_h = P_{hl} \times N_h = r \times N_h$.

Two-stage stratified probability proportionate to size designs

The formulae for the standard PIAAC selection probabilities for each stage are given below.

For the first-stage sample of primary sampling units (PSUs) in the remaining countries, let

m_h = number of PSUs to be sampled in stratum h;

MOS_{hi} = measure of size for PSU ii in stratum h; and
l_{psu}^{h} = sampling interval for the selection of PSUs in stratum h.

The probability of selecting PSU i in stratum h is

$$P_{hi} = \frac{m_h \, X \, MOS_{hi}}{\Sigma_{i \in h} MOS_{hi}} = \frac{MOS_{hi}}{l_{psu}^{h}}$$

For the second-stage sample of persons, let

n = total number of persons to be sampled;

N = total number of eligible persons;

n_{hi} = number of persons to be sampled in PSU i of stratum h; and

N_{hi} = number of eligible persons in PSU i of stratum h.

Let $r = n/N$, then the conditional probability of selecting person l in PSU i of stratum h is

$$P_{l \backslash hi} = \frac{r}{P_{hi}} \times \frac{l_{psu}^{h}}{MOS_{hi}}$$

The overall probability of selecting person l in PSU i of stratum h is

$$P_{hil} = P_{hi} \times P_{l \backslash hi} = r$$

The sample size in PSU i of stratum h is

$$n_{hi} = P_{l \backslash hi} \times N_{hi} = r \times \frac{\Sigma_{i \in h} MOS_{hi}}{m_h} \times \frac{N_{hi}}{MOS_{hi}} = r \times l_{psu}^{h} \times \frac{N_{hi}}{MOS_{hi}}$$

Three-stage stratified probability proportionate to size (PPS) designs

In a three-stage stratified PPS design, PSUs are selected with a probability proportionate to a measure of size as described below.

For PSU selection in the remaining countries, let

m_h = number of PSUs to be sampled in stratum h;

MOS_{hi} = measure of size for PSU ii in stratum h; and

l_{psu}^{h} = sampling interval for the selection of PSUs in stratum h.

The probability of selecting PSU ii in stratum h is

$$P_{hi} = \frac{\mathrm{M}_h \times MOS_{hi}}{\Sigma_{i \in h} MOS_{hi}} = \frac{MOS_{hi}}{l_{psu}^{h}}$$

For the second stage sample of dwelling units (DUs), let

d = total number of housing units to be sampled;

D = total number of housing units in the sampling frame;

d_{hi} = number of housing units to be sampled in PSU ii of stratum h; and

D_{hi} = number of housing units in PSU ii of stratum h.

Let $r = d/D$, then the conditional probability of selecting housing unit k from PSU i in stratum h is

$$P_{k|hi} = \frac{r}{P_{hi}} = r \times \frac{l_{psu}^{h}}{MOS_{hi}}$$

The overall probability of selecting housing unit k in PSU i of stratum h i

$$P_{hik} = P_{hi} \times P_{k|hi}$$

The DU sample size in a PSU is

$$d_{hi} = P_{k|hi} \times D_{hi} = r \times \frac{\Sigma_{i \in h} MOS_{hi}}{m_h} \times \frac{D_{hi}}{MOS_{hi}} = r \times l_{psu}^{h} \times \frac{D_{hi}}{MOS_{hi}}$$

For person selection, let

n_{hik} = number of persons to be sampled from housing unit k in PSU i of stratum h; and

N_{hik} = total number of eligible persons in housing unit k of PSU i in stratum h.

The conditional probability of selecting person l from housing unit k in PSU i of stratum h is

$$P_{l|hik} = \frac{n_{hik}}{N_{hik}}$$

The overall probability of selecting person l in housing unit k of PSU i of stratum h is

$$P_{hikl} = P_{hi} \times P_{k|hi} \times P_{l|hik} = r \times \frac{n_{hik}}{N_{hik}}$$

Sample units and sample selection methods

Sample units

Sample units may vary from country to country depending on the characteristics of the sampling frameworks available and on the number of sampling stages implemented. As a way of illustration, in Table 10 we show some examples from PIAAC countries that implemented a three-stage sample design. More details on the sample units used in different countries and options in this regard can be found in PIAAC technical documentation Mohadier et al. (2013).

Table 10. Examples of sample units for countries with three stages of sampling – PIAAC Round 3

Country	Stage 1	Stage 2	Stage 3
Ecuador	Cencus tracts	DUs	Persons
Kazakhstan	Localities	DUs	Persons
Peru	Area PSUs	DUs	Persons

Source: Mohadier et al. (2013)

Sample selection methods

Sample selection methods can also vary from country to country. As a way of illustration, in Table 11 we show some examples from PIAAC countries that implemented a two-stage sample design. More details on the sample selection methods used in different countries and options in this regard can be found in PIAAC technical documentation Mohadier et al. (2013).

Table 11. Examples of selection methods for countries with two stages of selection – Round 2

Country	Stage	Description
Israel (small localities)	1	Systematic PPS (number of persons aged 16-65 registered in the locality) from a sorted list within explicit strata
	2	Systematic random sample from a sorted list within explicit strata
Slovenia	1	Systematic PPS (number of persons living in the PSU)
	2	Systematic random from a sorted list

Source: Mohadier et al. (2013)

Stratification

Stratification combines sample units into homogeneous groups and reduces sampling variability between such groups and thus reduces the overall sampling variance associated with the resulting survey estimates. For this reason, when possible, stratification variables should be added to the sample selection process. However, to maximize the benefit of stratification, stratification variables should be reliable and related to the survey outcome. Countries with population registries available have the benefit of using person-level characteristics as stratification variables. Some examples of stratification variables used by countries participating in PIAAC are shown in

Table 12. Examples of stratification variables and methods for countries with two stages of selection – PIAAC Round 2

Country	Stage	Description
Israel (small localities)	1	Strata: combination of district or grouping of districts and type of locality. Within strata: Sort mainly by size of locality.
	2	Sort by geographic variables (district, type of locality, locality code, street code, and house number) and demographic characteristics (year of immigration and country of birth).
Slovenia	1	Sort by region and settlement type.
	2	Sort by settlement, street, house number, and surname.

Source: Mohadier et al. (2013)

Sample size determination

Adequate sample sizes are needed to establish stable item characteristics and to estimate separate population models for each tested language in a participating country. Population modelling is a critical step in obtaining appropriate proficiency values to be used in describing the distributions of skills in a country and in reporting national and subpopulation data.

The overall goal of the sample design is to obtain a nationally representative sample of the target population in each participating country that is proportional to the population across the country (i.e., a self-weighting sample design). Following the PIAAC design, the suggested minimum sample size requirement is 5,000 complete cases for the standard target population speaking the main language of the country.

Weighting

As mentioned before, a major objective of this assessment is to obtain accurate, precise, and internationally comparable estimates of population characteristics. Several considerations have to be taken into account to achieve this goal. This section describes the weighting procedures employed in PIAAC.

A final weight is required for all sampled persons with a completed instrument, and there are a number of steps in the development of the final weights intended for use in the estimation and analysis:

1. Assignment of a household base weight to each sampled household to compensate for differential probabilities of selection

2. Household-level eligibility and nonresponse adjustments to reduce potential biases arising from differences between respondents and nonrespondents

3. Assignment of a person base weight to each sampled person to compensate for differential probabilities of selection

4. Person-level eligibility adjustment and nonresponse adjustments

5. Trimming to reduce the impact of large weights, if necessary

6. Calibration of the person weights to independent control totals to compensate for noncoverage in the sample due to deficiencies in the sampling frame

The succeeding sections describe each of the weighting steps in detail. A summary of the adjustment factors and resulting weights at each weighting step is provided in Table 13.

Table 13. Adjustment factors and weights

Weighting step	Factor		Weight
Base weight	N/A		$W_1 = \frac{1}{P_{hl}}$
Unknown eligibility adjustment	$F_{1l} = \begin{cases} \frac{S_R + S_{NR} + S_{L1} + S_{L2} + S_D + S_I + S_U}{S_R + S_{NR} + S_{L1} + S_{L2} + S_D + S_I} \\ \frac{S_R + S_{NR} + S_{L1} + S_{L2} + S_D}{S_R + S_{NR} + S_{L1} + S_{L2} + S_D + S_I} \end{cases}$	If $l \in I$ If $l \in U$ If $l \in R, NR, L1, L2, D$	$W_1 F_{1l}$
Nonliteracy related nonresponse adjustment	$F_{3l} = \begin{cases} 1 \\ \frac{S_R + S_{NR} + S_D + S_U}{S_R} \\ 0 \end{cases}$	If $l \in L1, L2, I$ If $l \in R$ If $l \in NR, D, U$	$W_1 F_{1l} F_{3l}$
Literacy related nonresponse adjustment	$F_{4l} = \begin{cases} 1 \\ \frac{S_{L1} + S_{L2}}{S_{L1}} \\ 0 \end{cases}$	If $l \in R, I$ If $l \in L1$ If $l \in L2$	$W_1 F_{1l} F_{3l} F_{4l}$
Trimming	$F_{5l} = \begin{cases} 1 \\ \frac{cutoff}{W_1 F_{1l} F_{3l} F_{4l}} \end{cases}$	If $W_1 F_{1l} F_{3l} F_{4l} \leq cutoff$ If $W_1 F_{1l} F_{3l} F_{4l} > cutoff$	$W_1 F_{1l} F_{3l} F_{4l} F_{5l}$
Calibration	$F_{5l} = \frac{S^*}{S_R + S_{L1}}$ (for post-stratification) See Deming and Stephan (1940) for raking adjustments and Särndal, Swenson, and Wretman (2003) for GREG estimation.		$W_1 F_{1l} F_{3l} F_{4l} F_{5l} F_{6l}$

Source: Mohadier et al. (2013)

Preliminary steps in weighting

Countries should be responsible for selecting the variables that will be used in their nonresponse and calibration weighting adjustments. Prior to weighting, countries are required to evaluate the variables being considered for the weighting adjustments in their main sample.

For the nonresponse adjustment, variables need to be available for all eligible units and be related to proficiency and response propensity. The pool of potential nonresponse adjustment variables should come from the sampling frame or other external sources. A common source of nonresponse adjustment variables is normally a country census.

For the calibration adjustment, all variables selected by countries are required to have reliable control totals and be available for all respondents with age and gender collected. The quality of the data from the external sources had to exceed the quality of data from

the study itself (e.g., the mean square errors of the external estimates needed to be smaller than those of the uncalibrated estimates from the survey). The concepts, definitions and coverage of the data (counts) from the external sources need to be the same as those employed by this study. Additionally, the year of the control totals needed to be as close to the data collection period as possible, ideally covering the same time period as the field period. Variables used for nonresponse adjustment and in calibration must have less than 5% missing data.

Household-level weighting adjustments

This section outlines the weighting process at the household level, which included the creation of the household base weights that reflected the household selection probability and was adjusted for unknown eligibility and nonresponse.

Household base weights

The household base weight is assigned to all sampled households and is computed as the reciprocal of the household selection probability. For countries with a multistage sample design, the household selection probability should correspond to the product of the conditional selection probabilities at each stage. For example, if households are selected within primary sampling units (PSUs), then the household base weight would be

$$W_k = \frac{1}{P_{hi}P_{k|hi}}$$

where P_{hi} is the probability of selecting PSU i in stratum h, and $P_{k|hi}$ is the conditional probability of selecting household k within PSU i of stratum h.

The household selection probability also reflects any duplicate records in the sampling frame or any changes to the subsampling procedures.

Household unknown eligibility adjustment

Before any household-level nonresponse adjustment is applied, an adjustment for unknown eligibility should be performed if the eligibility status of some households cannot be determined. In this step, a portion of the weights of the households with unknown eligibility status (i.e., whether they contained a person age 16 to 65) is distributed to ineligible cases. An adjustment factor is computed as the proportion eligible among those with known eligibility status to down-weight the cases with unknown eligibility status (accounting for an estimated proportion that was ineligible). The down-weighted unknown eligibility cases are then treated as eligible nonrespondents. This adjustment is done within weighting cells defined for the unknown eligibility adjustment.

Household nonresponse adjustment

For the nonresponse adjustment, the nonrespondents have to be divided into two categories. The first consists of cases involving non-literacy-related nonresponse. Examples of this category include refusals and nonresponse due to speech impairment. Non-literacy-related nonrespondents are likely to be similar to respondents with respect to proficiency scores. The second category is literacy-related nonresponse. Language problems should be the only type of literacy-related nonresponse at this level. Households with this type of nonresponse are presumed to differ from responding households with respect to proficiency. Therefore, the weighting procedures should adjust the weights of the respondents to represent the non-literacy-related nonrespondents only. The contribution of the literacy-related nonresponse to the total population is accounted for by the literacy-related nonresponse adjustment carried out at the person level.

The next step in the weighting process is to adjust the unknown eligibility-adjusted weights to reduce potential bias as a result of nonresponse. An adjustment is made to distribute the unknown eligibility-adjusted weights of the non-literacy-related nonrespondents to the respondents. The nonresponse adjustment was performed within cells that were defined based on pre-selected weighting variables that were found to be related to proficiency and to response propensity. Within each adjustment cell, the household unknown eligibility-adjusted weights of nonrespondents are redistributed over a relatively large pool of cases (approximately 30 or more respondents). Additionally, the amount of variation in the nonresponse adjustment factors should be kept to a minimum by limiting the maximum allowable nonresponse adjustment factor, which is a function of the achieved response rate.

Person-level weighting adjustments

This section describes the process of creating the person-level weights, including the computation of person base weights; the person unknown eligibility adjustment; the nonresponse adjustment procedure designed to reduce potential nonresponse bias; the calibration of weights to control totals; and the general trimming procedure used to reduce the impact of extreme weights.

Person base weights

The person base weights account for both nonresponse to the household and differential within-household selection rates. The person base weights are computed as the product of the household nonresponse-adjusted weight and the reciprocal of the within-household person selection probability. Ft

The base weight for each sampled person is computed as the reciprocal of the person selection probability.

Person unknown eligibility adjustment

An adjustment for person unknown eligibility should be performed if the eligibility status of some sampled persons cannot be determined due to the inability of the survey to locate and interview these selected persons not residing at the address listed in the registry. In the person unknown eligibility adjustment, a portion of the person base weights of the sampled persons with unknown eligibility status is distributed to the ineligible cases. An adjustment factor is computed as the proportion eligible among those with known eligibility status to down weight the cases with unknown eligibility status (accounting for an estimated proportion that was ineligible). The down-weighted unknown eligibility cases are then treated as eligible nonrespondents in the nonresponse adjustment.

Person non-literacy-related nonresponse adjustment

For the nonresponse adjustment, the nonrespondents are divided into two categories. The first category consisted of non-literacy-related nonrespondents (e.g., refusals and inaccessibles with known eligibility) and sampled persons with a disability (e.g., hearing impairment and physical disability). They are likely to be similar to respondents with respect to proficiency scores. The second category is literacy-related nonresponse. Types of literacy-related non-response include language problems, reading and writing difficulty, and learning mental disability. Sampled persons with this type of nonresponse are presumed to differ from respondents with respect to proficiency. Therefore, literacy-related nonresponses receive a different treatment than non-literacy-related nonrespondents.

As mentioned earlier, an adjustment is made to distribute the person base weights of the non-literacy-related non-respondents and sampled persons with a disability to the respondents' weights.

Excluded inaccessible sampled persons are treated as non-literacy-related non-respondents in weighting. An adjustment is made to distribute the person unknown eligibility-adjusted weights of the non-literacy-related non-respondents, sampled persons with a disability, and down-weighted unknown eligibility cases to respondents.

The nonresponse adjustment is performed within cells that are defined based on preselected weighting variables that are found to be related to proficiency and to response propensity. Within each adjustment cell, the person unknown eligibility-adjusted weights of nonrespondents are redistributed over a relatively large pool of cases (approximately 30 or more respondents). Additionally, the amount of variation in the

nonresponse adjustment factors is kept to a minimum by limiting the maximum allowable nonresponse adjustment factor, which depended on the achieved response rate.

Person literacy-related nonresponse adjustment

Person literacy-related weights are also adjusted by non-response. This adjustment is necessary primarily to allow literacy-related nonresponse to be represented in the calibration procedure. This adjustment assumed that all types of literacy-related nonresponses were similar in proficiency.

Calibration

To address undercoverage bias, to reduce the mean square error of estimates and to create consistency with statistics from other studies, the next weighting step is to adjust the survey weights to match population control totals. At a minimum, weights should be benchmarked to control totals for age and gender. Respondents who completed the instrument should receive a final weight and be included in the calibration.

Three main calibration techniques normally employed by countries are post-stratification, raking and generalized regression estimators (GREG). Post-stratification adjusts survey weights of respondents so that the weighted sample distribution is the same as some known population distribution (i.e., the sums of the adjusted weights of the respondents are equal to known population totals for certain subgroups of the population). The raking procedure uses an iterative procedure to adjust the survey estimates to the known marginal totals of several categorical variables. The GREG estimator is a model-assisted approach that can be used to adjust weights to exploit explicitly the relationship between a survey variable and auxiliary variables. It is suggested that the techniques that best adapt to the needs and context of the country is used in this step.

Trimming the outliers

Even a carefully designed sample cannot fully prevent the need for reducing extreme weights. Sample designs that included the selection of dwelling units normally have more variability in the weights compared to directly sampling persons from registries because of unequal household sizes. The use of nonresponse and calibration adjustments also introduced variations in sampling weights. Weight trimming introduced some bias into the sampling weights. However, the trimming adjustment in most cases reduced the sampling error component of the overall mean square error more than it increased the bias as the adjustment was applied to only a relatively small number of weights (Lee, 1995).

The person weights should be trimmed as necessary after the first calibration. Using a design-based procedure, cells for trimming are formed from groups that were expected to be approximately self-weighting. In each cell, weights above a cutoff value are trimmed down to the designated cutoff. To define the trimming cut point, the coefficient of variation (CV) can be used based on the weights after raking. In PIAAC, weights were trimmed when they were over $3.5 \times \sqrt{1 + CV^2}$ times the median raked weight (within each trimming cell, if sampling rates varied by sampling domains). During trimming, the trimming factor is applied to each replicate weight. After trimming, the weights should be recalibrated back to the control totals.

Variance estimation

Inferences will not be valid unless the corresponding variance estimators appropriately reflect all of the complex features of the proposed sample design (e.g., stratification and clustering). The replication approach is used for estimating variances for the international analyses of PIAAC data. Under the replication approach, subsamples (also known as replicates) from the full sample are formed and statistics of the subsamples are used to estimate the variance of the full sample statistic. The replication approach, in conjunction with the multiple imputation approach used to derive the plausible values, captures the variation due to the complex sampling and estimation approaches, including:

- Sample design
- Selection
- Weighting adjustments
- Measurement error through the processing of multiple imputation of plausible values

Replication methods are applied to surveys by dividing the sample into specially designed replicate subsamples that mirror the design of the full sample. To form the replicate subsamples, variance strata and variance units are defined. Each subsample is reweighted to account for the subsampling that occurred. An estimate is then calculated for the full sample and each of the replicate subsamples. The variance of the full sample estimate is computed as the sum of squared deviations between each replicate subsample estimate and the full sample estimate. The general replication formula is

$$Var(\hat{\theta}) = c \sum_i (\hat{\theta}_i - \hat{\theta}_0)^2$$

where

c = 1, for the paired jackknife (JK2)

= (g-1)/g, for the random groups (delete-one) approach (JK1)

= 1 / g for the BRR approach

= 1/[g(1-k)2] for Fay's method

g = number of replicates

k = weighting factor for Fay's method

$\hat{\theta}_0$ = full sample estimate

$\hat{\theta}_i$ = estimate for replicate i

A variety of sample designs were employed across the different countries participating in PIAAC. Replication is adaptable to a wide variety of designs, including simple random sampling, systematic sampling, stratified designs and multistage cluster designs. In general, replication schemes are selected based on the sample design. A random groups approach may do well for a simple random sample while a paired jackknife mechanism is not meant for an SRS, but could be adapted. The paired jackknife would work very well for a one-PSU per stratum design, while a random groups design is not appropriate. Some efficiency is gained by selecting the most appropriate approach for the sample design.

Creation of replicate weights

The specification of variance strata and variance units must conform to the design assumptions of a replication method and should be determined by the type of sampling design that was used to collect the data (e.g., whether or not stratification was used and how many PSUs were in each stratum). In addition, in some cases, the sampling strata and PSUs may need to be grouped to reduce the number of replicates to fit the sample design into a replication design that follows the PIAAC standards.

Once the variance strata and variance units were assigned, replicate base weights are created. The household or person base weights are replicated. Subsequently, all weight adjustments that were conducted for the full sample have to be conducted on each replicate weight to capture the variation created, or reduced, by the weight adjustments.

Accounting for imputation error variance component

For estimation using plausible values (PVs), calculations must account for both the sampling error component and the variance due to imputation of proficiency scores. The estimator of the population mean is the average of the *M* PV means,

$$\hat{Y}^* = \sum_{m-1}^{M} \hat{Y}_m / M$$

The variance of the estimated mean $\hat{Y}^*$ is computed using formulas specific to PVs as follows:

$$v(\hat{Y}^*) = U^* + B\left(1 + \frac{1}{M}\right)$$

where, the "within" variance component is computed as the average of the sampling variance for each of the *M* plausible values, computed as,

$$U^* = \left(\sum_{m-1}^{M} U_m\right)/M$$

where the sampling variance of the estimated mean $\hat{Y}_m$ for plausible value m is U_m, and where, the "between" component is calculated as

$$B = \left[\sum_{m-1}^{M} \left(\hat{Y}_m - \hat{Y}^*\right)^2\right]/(M-1)$$

where, the mean of each of the *M* PVs γ_{l1}, γ_{l2},..., γ_{lM} for sample unit l is computed as

$$\hat{Y}_m = \frac{\sum_{l \in S} w_l \, y_{lm}}{\sum_{l \in S} w_l}; m = 1, \ldots, M$$

where S denotes the set of sample units.

The standard error is computed as the square root of the total variance, $\sqrt{v(\hat{Y}^*)}$

Specialised software

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 does not always include procedures for replication techniques, however, there are several pieces of software that have been specially developed for these kinds of statistical procedures. Below, there are some examples of different pieces of software that are well documented and which documentation includes examples and exercises:

IEA IDB Analyzer

IEA IDB Analyzer (IEA, 2019) is a plug-in for the Statistical Package for the Social Sciences (IBM, 2015) and SAS (SAS, 2012) that allows the user to combine and analyze data from

IEA's large-scale assessments. The application can be downloaded at https://www.iea.nl/data-tools/tools

Replicates

Replicates (ACER, 2018) is an add-in component running under SPSS and offers a number of features for applying different replication methods when estimating sampling and imputation variance. The application can be downloaded from https://iccs.acer.org/ICCS2016reports

WesVar

WesVar (Brick et al., 2000) is a computer programme developed by Westat that allow users to compute estimates, replicate variance estimates, and to import and export data to creating weights, generating statistics, and obtaining regression output with survey data with complex sample and assessment designs. The application can be downloaded from
https://www.westat.com/capability/information-technology/wesvar

Intsvy (R)

Intsvy (Caro & Biecek, 2017) is an R package that provides tools for importing, merging, analysing and visualizing data from international assessment studies (TIMSS, PIRLS, PISA, ICILS, and PIAAC). It can be downloaded at https://cran.r-project.org/web/packages/intsvy/index.html Learning resources and video tutorials can be found at https://www.youtube.com/channel/UCyykJxYbj_WGIZH5AttwyjQ

RALSA (R)

The R Analyzer for Large-Scale Assessments (RALSA) (Mirazchiyski, 2021) is an R package for preparation and analysis of data from large-scale assessments and surveys which use complex sampling and assessment design. RALSA is a free of charge and open-source software, it works on any system which can run a full installation of R. In addition to the traditional command-line R interface, RALSA has a Graphical User Interface that can be used in any web browser. The user guide and learning materials can be accessed at http://ralsa.ineri.org/
Dumais & Gough (2012) also prepared a complete series of examples and exercises, including example data and software routines, to estimate replicate weights Jacknihe variance estimation, as well as how to calculate mean differences while considering a complex sample design. The exercises, 14.2, 14.4 and 14.6 are particularly relevant for the methods used in this assessment.

7. Logistic of the national assessment

The coordination of national logistics determines, to a large extent, the success of the assessment. The potential needs of the staff, the procedures for contacting schools, the availability of facilities, and the distribution of the instruments are all relevant issues to ensure. Most of the information included in this chapter is dapted from Howie and Acana (2012).

Staff recommendation and contacting schools process.

Considering that national assessment is to provide valid information about educational achievement or the opinions of the students in the target population about specific topics, the decisions regarding the personnel who will carry out the assessment and the facilities they will need are crucial. All sorts of problems can be anticipated if personnel are not competent or if facilities are inadequate.

As a general principle, not only should personnel have specialist skills; they should also be committed and open-minded, attentive to detail, and willing to put in additional hours beyond the normal workday. From the point of view of technical adequacy and efficiency, these attributes are more important than seniority within a government department or within an academic institution.

This section describes the role of typical staff members[8] (for example, the national coordinator) as well as the roles of additional personnel, such as test administrators, who will be required to carry out the assessment. A list of the personnel considered here and a description of their main functions are in Table 14.

Table 14. Main staff members involved in the logistics of the assessment and their functions

Staff member	Description and main functions
National research coordinator	Should give general direction and provide leadership throughout the planning and implementation stages of the national assessment. Should be respected within the educational community, should have access to key educational stakeholders and to the main sources of funding, should be familiar with concepts in education and measurement. He or she should be able to see the "big picture."

[8] There are other staff members that could be involved at different stages, for example, item writers, test administrators, statisticians, data managers, designers, translators, data entry personnel, data recorders, tests scorers, among others.

Staff member	Description and main functions
Assistant National Coordinator	May be required depending on the structure of the education system, the scope of the assessment, the time demands on the NC, and the availability of funding. The assistant NC should have many of the attributes required of the NC and should support and serve as a substitute for the NC when necessary.
Regional Coordinator	In large countries with regional administrative systems, the national assessment team should consider appointing regional coordinators to organize testing and to liaise with schools and test administrators. Such coordinators would be responsible for allocating and delivering materials to the test administrators and should check the contents of boxes coming from the central office.
School liaison person	The school liaison person or school coordinator could be a teacher or guidance counselor in a school, but he or she should not be teaching students selected for the assessment. Frequently, the school principal serves in this role. The school liaison person serves as a contact point in schools for the national assessment team and helps ensure that school personnel are aware of the assessment. This staff member is the key for coordination with administrators and other participants, such as parents or teacher (when they participate).
Test administrators	Include the distribution of the student test instruments according to the student tracking forms, the supervision of test sessions, ensuring that the timing of the test sessions was correct, and recording student participation. in some countries, classroom teachers administer national assessment tests to their own students. More often than not, however, teachers other than those who teach the students who are taking the test or individuals who are external to the school are entrusted with this task. In some countries, data collection is contracted to a body that specializes in that activity. Potential administrators should have the following characteristics: • Good organizational and communication skills • Experience working in schools • Reliability, and ability and willingness to follow instructions precisely

Source: Adapted from Howie and Acana (2012)

The national coordinator should inform schools that they have been selected for the national assessment as soon as possible when its are selected[9]. If required, the permission of the ministry of education or regional education authority should be obtained before schools are contacted. When schools are contacted and invited to participate, they should be asked to acknowledge receipt of the invitation. The school should be asked to appoint a contact person, school liaison person, or coordinator for the assessment. The national assessment team should strive to ensure that it establishes and maintains a good rapport with local education authorities, if they exist. The national assessment team should keep an updated list or tracking form of participating schools to help monitor fieldwork progress. The form will provide information on schools, such as school name, size, and contact information (see Table 15 as an example).

Additionally, there several facilities that are relevant in the administration of a national assessment. This aspect refers to the space, equipment and or tools for staff members. For instance, space for meetings, access to rooms, space for organizing and storing materials, technological tools for different activities involved in the assessment (phones, computers, internet, software's, son and so far).

Table 15. Example of a National Assessment: School Tracking Form

Priority of School[a]	School ID	Name, address, phone number of school	Name and phone number of school coordinator	School size	Status (participant or non-participant)	Date materials sent	Date materials received	Date of testing
1								
1								
1								
1								
1								
2								
2								
2								
2								
2								

a. Schools selected from the sample are priority 1. Replacement schools are priority 2.

[9] Insofar as possible, after schools have been selected, they should not be changed or replaced. Despite the best efforts of a national assessment team, however, some school replacements may be necessary. Should the need to replace schools be anticipated, that possibility should be discussed with the sampling statistician so that adequate sampling procedures are implemented, and replacement schools are properly selected. Under no circumstances should the selection of replacement schools be left to the discretion of the test administrator or local school official.

Logistic in instrument checks and distribution

The national coordinator or his or her appointee should check the quality of all tests, questionnaires, and manuals to ensure the following:

- Spelling and typographical errors are removed.
- Font size in test booklets is sufficiently large. Large font sizes are particularly important for young children.
- Adequate spacing is used between lines of text.
- Diagrams are simple and clear. Where possible, they should be on the same page as the relevant text.

A qualified data entry person who is familiar with computer packages such as Microsoft Office should type tests, questionnaires, and other materials. Likewise, cost-saving measures that should be considered at this stage include the following:

- Preparing test booklets to fit on an even number of pages
- Careful proofreading, especially of final drafts, which can help prevent reprinting of test booklets necessitated by serious typographical or graphical errors
- Giving the printer adequate time to print tests and questionnaires to avoid paying overtime rates when the assignment has to be completed over a relatively short time or when the printer has other priorities

At least, three people should independently proofread final drafts of all the materials used in a national assessment. When print runs are ordered, additional copies should be requested for each school package in anticipation of the need for replacement schools and of some spoilage.

Effective national assessment team leaders plan thoroughly and well in advance of administration of the assessment in schools. They also tend to delegate responsibility while retaining overall control of the preparation process through quality control measures, in particular spot-checking the work of others.

A set of packing procedures should be established and documented. A packing checklist is required. National assessment staff members should sign and date the appropriate boxes in the “Packed” and “Returned” columns in the packing check- list. The school liaison person is expected to do the same in the boxes in the “Received” columns after checking the material sent from the national assessment office.

Local circumstances will determine the most appropriate and cost- effective method of delivering and collecting materials for the national assessment. In some instances, materials are delivered to central offices that are secure (for example, district education

or local government offices), and test administrators collect them using public transportation. In other cases, where secure and reliable delivery systems exist, materials are delivered to test administrators' homes. Sometimes, teams of administrators travel together in a van and are dropped off with the necessary materials at schools.

In some national assessments, test administration is carried out at the same time in all schools, usually over one or two days. In others, test administrators travel from school to school over a short period. In the latter case, care has to be taken to maintain the security of test materials and to ensure that test-related information is not exchanged between schools.

Administration and common problems

Problems associated with administering a national assessment tend to vary from country to country in both nature and magnitude. The more serious the problem, the more it undermines the entire national assessment enterprise. From the outset, the national assessment team should ensure that the sampled schools are in fact the ones in which students are being assessed. Some teams have discovered "ghost" (bogus) schools after using national data sources for sampling purposes. The test administrator and the school liaison person should establish that the pupils who take the tests are in fact the pupils who were selected for participation.

The following are other problems that have been identified in administration:

- Date of testing clashing with a school event.
- Pupils completing the first section of the test and leaving school before the second section.
- Teachers and pupils arriving late.
- Teachers, and even the principal, insisting on remaining in the class while students are taking the test.
- Lack of adequate seating arrangements for test taking.
- Failure to stick to time limits.
- Test administrator or others giving assistance to students.
- Copying by pupils.

High levels of participation are required in a national assessment to provide valid information on student achievement in the education system. IEA studies, for example, require a participation rate of at least 85 percent for both schools and students or a combined rate (the product of school and student participation) of 75 percent. IEA also sets the upper limit of exclusions (on grounds such as school remoteness and disability) at 5 percent of the desired target population (see Chapter 6 for more details on participation rates). In an effort to improve the level of school cooperation, replacement

sessions could be held at a later date for students who were absent for the initial assessment session. Experience suggested that students and schools tend to cooperate more fully when they realize that the test administrators would keep returning until all selected pupils have been tested.

Quality issues

To monitor the quality of test administration, the test administrator should complete a test or questionnaire administration form (Figure 4) after work in an individual school has been completed. The form will provide a record of the extent to which proper administrative procedures were followed.

Figure 4. Example ofa Test Administration Form

Complete one form per testing session.

Name of test administrator: _____

School ID: _____

School name: _____

Class name: _____

School liaison person: _____

Original testing session: _____

Replacement testing session (if applicable): _____

Date of testing: _____

Time of testing

Start time	End time	Details
		Administration of test materials
		Testing session 1
		Testing session 2
		Testing session 3
		Testing session 4

Source: Howie and Acana (2012)

To check further if testing has been carried out following prescribed procedures, many national assessments appoint a small number of quality control monitors to make unannounced visits to schools. Although all test administrators should know that a

possibility exists that they will be monitored, in practice, usually only 10 to 20 percent of schools are visited. Quality control personnel should be familiar with the purpose of the national assessment, the sampling design and its significance, the roles of the school coordinator and test administrator, the content of tests and questionnaires, and the classroom observation record. They should be briefed on how to conduct school visits without disrupting the actual assessment. Monitors should complete a form on administrative and other conditions in each school visited. Examples of the activities for which information is recorded in the form used for TIMSS (Trends in International Mathematics and Science Study) are provided in .

Figure 5. Examples of Questions Addressed by Quality Control Monitors in TIMSS

1. Preliminary activities of the test administrator

Did the test administrator verify adequate supplies of test booklets? Were all the seals intact on the test booklets prior to distribution?

Was there adequate seating space for the students to work without distraction?

Did the administrator have a stopwatch or timer?

Did the test administrator have an adequate supply of pencils and other materials?

2. Test session activities

Did the test administrator follow the test administrator's script exactly in (a) preparing the students, (b) distributing materials, and (c) beginning testing?

Did the test administrator record attendance correctly? Did testing time equal the time allowed?

Did the test administrator collect test booklets one at a time from the students?

3. General impressions

During the testing session, did the test administrator walk around the room to ensure that students were working on the correct section of the test and behaving properly?

In your opinion, did the test administrator address students' questions appropriately?

Did you see any evidence of students attempting to cheat on the tests (for example, by copying from a neighbor)?

4. Interview with the school coordinator

Did you receive the correct shipment of items?
Was the national coordinator responsive to your questions or concerns?

Were you able to collect completed teacher questionnaires before test administration?

Were you satisfied with the accommodation (testing room) for the testing? Do you anticipate that makeup sessions will be required at your school?

Did students receive any special instruction, motivational talk, or incentive to prepare them for the assessment?

Were students given any opportunity to practice questions like those in the test before the testing session?

Source: Howie and Acana (2012)

8. Data preparation

In this chapter, we refer to data preparation for all the steps required from the data entry process to the generation of the data release for inquiry. Here, the objective is to minimize any possible error that may distort the collected responses when these are stored in digital format for further use (Brese & Cockle, 2017).

Data cleaning

Data cleaning encompasses all data related process from data importation to the data release. The purposes of these different tasks are to turn the raw data from the collected responses, into useable data files for inquiry. Brese & Cockle (Brese & Cockle, 2017) enlist the following common steps implemented in large scale studies:

- Import data
- Structure Checks
- Values Ranges
- Id checks
- Linkage checks
- Background checks
- Merge scores and weights
- Export

Import data. Import data refers to the process of taking the files generated during the data entry process and turn these into actionable files within a statistical software (e.g. SAS, SPSS, STATA, R). In studies where data collection occurs via a web platform, or other forms of software, instead of a paper-based survey, responses do not come from a data entry process. Yet, the generated data files from these applications would still need to be imported to a statistical software environment to proceed with the data cleaning process. As such, data importation is the step where raw data that contains participants responses and measures are turned into analyzable files.

***Structure checks*.** These checks refer to the structural features of the expected data. For example, the received data should conform to available *codeplans*. These codeplans are brief documents that are used during the data entry process. These documents specified how responses are coded by data clerks, to "entry" participants responses to the instrument using certain values. In these codeplans all coded responses are enlisted. Thus, the imported file should have a specific number of columns that represent each expected variable. A common problem during structure checks is the importation of data that contains text field or text strings. Most of the standard files format separate data fields (i.e., variables) using spaces, tabs, "," or ";", representing different columns. However, if typed responses contain spaces, "," or ";" data importation may incur in errors, by miss-representing the expected columns per response. Structural problems during data importation might be spotted in the structure checks phase.

***Values ranges*.** Following the codeplans, all collected responses after a study should have a specific range of valid values. Any other value outside these ranges could be deemed invalid, following an agreed codeplan between the data entry step and the data

preparation step. During these checks, is expected that the data entry is the result of a systematic and documented process. Thus, for example, if the data entry process is managed by two different data centres or data entry teams, these should have followed the same codeplan. In essence, that each team typed the same value, for the same response, over the same item and questions. During the values ranges check, any deviation should be identified, amended and documented. Numeric typos attributed to data clerks' errors are expected to be identified in this step. Systematic errors attributed to software features, from studies using software applications to collect responses are also expected to be picked up in this step.

***Id checks*.** Single application studies assume a participant provides answers only once in a study. Thus, a common convention is that participants appear in a data record only once, and no participant id can be repeated. *Id checks* consist of making sure the previous convention is fulfilled. A common scenario in studies where paper and online participation is open for participants, a participant could appear twice in the raw data response records (Brese & Cockle, 2017). In these cases, one of the records should be selected, and document which one was selected (e.g. the earliest participation), and avoid the unnecessary duplication of a case.

***Linkage checks*.** Multi-actor studies include different participants related to each other by some structure. In the case of large-scale studies in education, the most common example of these relations is the linkage of the school principal, teachers, and students to their respective school. The Linkage checks refer to the process of assuring all linkages are complete, consistent and logically correct. This process assures that information from different sources, including participant responses and other records can be put together into an analyzable data table.

***Background checks*.** This step refers to assuring the consistency of information from participants. For example, a student may give their age and sex in a context questionnaire in the study. However, the same study may have sociodemographic records from all participants where the same information is also present. During this step of the data cleaning, is possible to opt the information retrieved using the sociodemographic records if these are deemed more reliable. Likewise, if answers from two different questions should present certain consistency, this expected consistency can be evaluated and amended if necessary. For example, an immigrant student could indicate his/her age in one question and his/her number of years in the country in another question. The second typed response should be a smaller numeric value than the first. During this stage is expected these inconsistencies are resolved by clarifying if these are the result of data entry error, a typo from the participant. Moreover, during this stage, if this inconsistency

is not resolved and kept in the data file, documentation should be provided so users of the release data know these were allowed.

Merge scores and weights. Large scale studies often include the preparation of survey weights, and the generation of scores to summarize responses to test and scales. These types of data are often handled separately from participants responses. In this step, these records are added to the data response file. In this step any unexpected inconsistency between the list of cases with survey design, the list if cases with scores and the list of cases with responses should be identified and documented. For example, a student may present valid responses to all instruments. Yet, the school which the student is a member of may have been dropped from the study due to the low rate of participation. As such, the student record doesn't present survey weights or scores. Thus, during the merge of records, is expected to establish what is the valid list of cases of the study, and what records are discarded, if any.

***Export*.** Release data is generated at this stage, containing only variables for inquiry. Any other variable generated during the data cleaning process is erased or removed.

Data cleaning steps might be done iteratively (Brese & Cockle, 2017) till the expected data consistency is reached. In summary, the data cleaning process includes all the actions necessary to turn the raw data of collected responses into analyzable data files. Additionally, this process also includes the task of amending or excluding participants records not sufficiently consistent or reliable for the purpose of the study. Thus, the data cleaning process also consists of establishing what are the valid responses for further use and establishing what is the valid list of cases for further use.
Is recommended all data cleaning should be conducted following a reproducible process. This enables to ensure that all changes to the raw data are documented, and repeated if necessary. In practice, such a process can be implemented using reproducible research practices and literal programming in a statistical environment (Baumer et al., 2014).

Codebook

What is a codebook?

Codebooks are technical documentation that allows user to interpret stored data. Codebooks should accompany data files, so the stored values can be used to import data in statistical software, produce interpretable results such as descriptive analysis, and model-based results. In essence, these documents act as a dictionary regarding what a value in a data file means. As such, an exhaustive codebook can have as many entries as variables a data file has (Gebhardt & Berezner, 2017). In general, in large scale assessment

studies this documentation can be found in three different sources: codebooks are partially embedded in the release public data files, in the technical report from the study and in its user guide. The relevance of codebooks lies in their role of conveying information for the interpretation of the stored data.

Codebooks are generated before data is collected, and when data is released for inquiry. When responses are being collected a codebook allows users to match items and participants responses. This documentation allows distinguishing between expected values, and non-expected values, thus aiding data validation and data cleaning procedures. For example, if a question presents a response space of two categories in the final application, coded as 1 and 2 in the data entry process, then all values different from the coded values can be deemed invalid to represent participant responses (Gebhardt & Berezner, 2017). Codebooks from the production stage may contain more coded events than participants responses, such as “not reached” and “not administered” items, containing process information (Provasnik, 2021). This type of documentation, generated during the production stage and data entry are also called “codeplans” (Brese & Cockle, 2017). These documents purpose is to aid the data entry process. In contrast, codebooks for released data may contain a selection of the coded values. That is, it may contain only the coded values for each valid participant response while excluding other coded events used in the data validation and data cleaning process. The present section of this guideline is focused on codebooks for data files released for inquiry.

Types of codebook

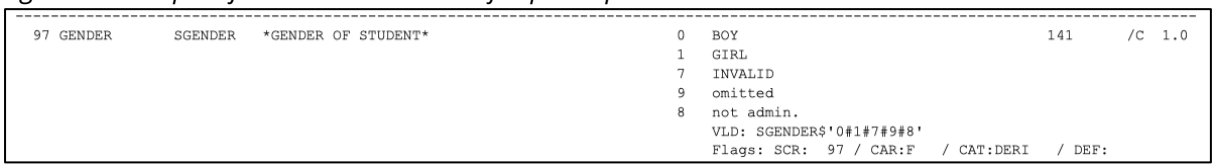
Codebooks are built in different formats and styles. Some codebooks are very succinct, containing just enough information regarding what variables constitute an indicator. Other codebooks are much more detailed, including how original responses are recoded to generate an interpretable score in a certain way. A different style of codebooks is instrument embedded codebooks. These codebooks contain less information regarding how original responses are recoded yet are very explicit regarding the instrument the participants interacts with to produce responses. And finally, codebook documentation can be presented as data file embedded codebooks. This latter type of codebook contains similar information than its previous counterparts but is stored in the release data file from the study. In the following section, we include examples of these different types of codebooks.

A simple example of these different types of codebooks can be illustrated using participants’ sex. Participants’ sex is often coded with two values: one and two. To register participants sex, they get asked a closed-form question with a two-option response space. The following figures are examples of how participants sex is documented in a succinct codebook, data file embedded codebook, detailed codebook, and with an

instrument embedded codebook. For illustration purposes, we will use participants sex from ICCS 2016 study.

What we are calling succinct codebooks are often generated using statistical programs (e.g., SAS, SPSS, STATA), and will consists of a table that include the names of the variable, the label of a variable, and its response values.

Figure 6. Example of a succinct codebook for participants sex indicator



```
 97 GENDER       SGENDER   *GENDER OF STUDENT*                    0   BOY                                  141      /C  1.0
                                                                  1   GIRL
                                                                  7   INVALID
                                                                  9   omitted
                                                                  8   not admin.
                                                                      VLD: SGENDER$'0#1#7#9#8'
                                                                      Flags: SCR:  97 / CAR:F   / CAT:DERI   / DEF:
```

Source: ICCS 2009 public data file (Köhler et al., 2018, p. 276)

Embedded codebooks are metadata that's comes in the study data files. To access this metadata, data files need to be open in statistical software (e.g. SAS, SPSS, STATA, R) that handles labelled vectors. That is software that can read and embed metadata onto data tables. The following example corresponds to an output in R, to get codebook documentation of participants' sex from ICCS 2016 data files.

Figure 7. Example of data file embedded codebook for participants sex indicator displayed in R

```
> # variable table
> data_iccs %>%
+ dplyr::select(S_GENDER) %>%
+ r4sda::variables_table() %>%
+ knitr:::kable()


|variable |type    |values                                          |labels         |
|:--------|:-------|:-----------------------------------------------|:--------------|
|S_GENDER |dbl+lbl |, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1,… |Student gender |
>
> # variable label
> r4sda::variable_label(data_iccs$S_GENDER)
[1] "Student gender"
>
> # value labels
> r4sda::value_labels(data_iccs$S_GENDER)
# A tibble: 5 x 2
  value label
  <chr> <chr>
1 0     Boy
2 1     Girl
3 7     Invalid
4 8     Not administered
5 9     Omitted
```

Source: ICCS 2016 public data, see https://www.iea.nl/index.php/data-tools/repository/iccs

The detailed codebook contains the same information as the previous formats, while also including the primary question, from which the variable is generated, and the operation used to create it.

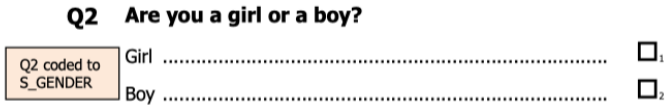
Figure 8. Example of a detailed codebook for participants' sex indicator

Variable Name	S_GENDER		
Description	Student gender		
Procedure	Simple recoding		
Source	Are you a girl or a boy?	IS3G02	Recoding
	Girl	1	1
	Boy	2	0

Source: International student questionnaire from ICCS 2016 (Köhler et al., 2018, p. 276)

Finally, the instrument embedded codebook includes the values for each response, overlayed on top of a representation of the instrument the participants interact with to generate their responses.

Figure 9. Example of an instrument embedded codebook for participants sex indicator



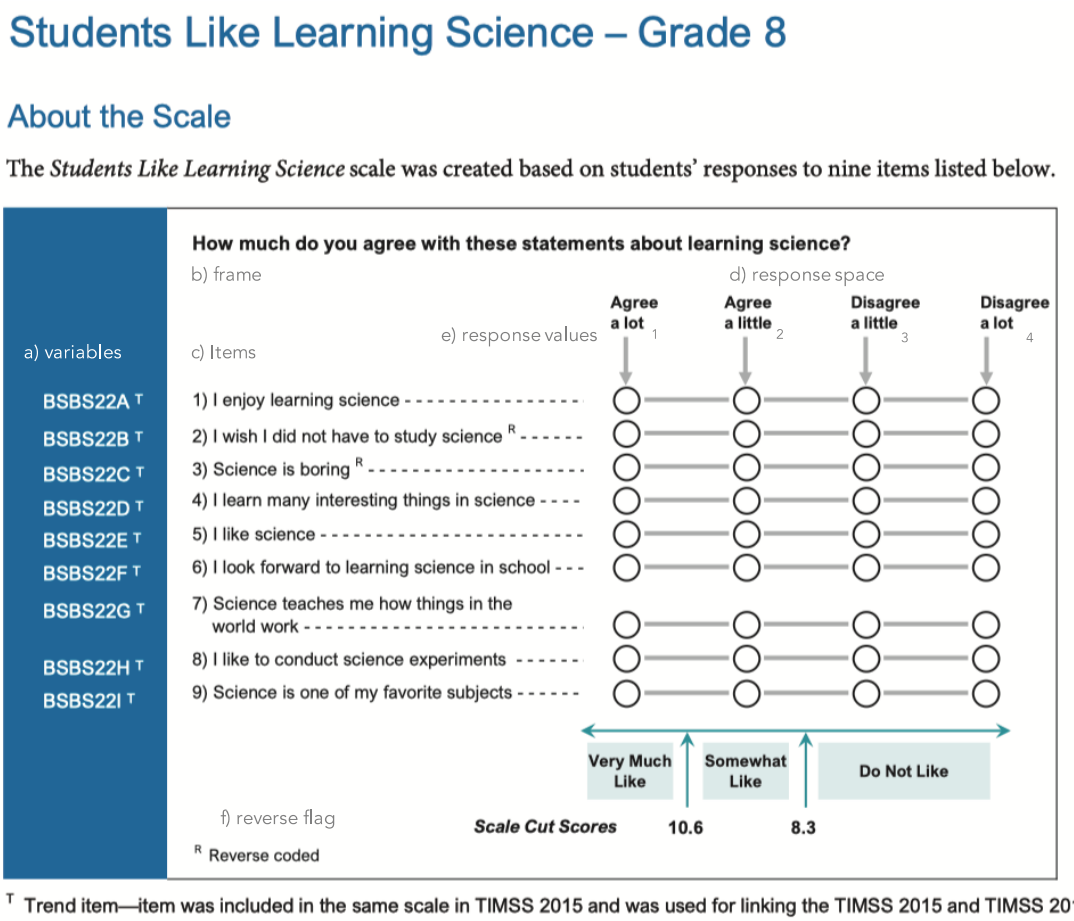
Source: International student questionnaire from ICCS 2019 (Köhler et al., 2018)

Elements of a codebook

To illustrate the main elements of a codebook we use "Students Like Science" scale from TIMSS 2019 Technical report (Yin & Fishbein, 2020, p. 16.259). In particular, its instrument embedded codebook. In the following figure, we highlight the elements of interest: a) the names of the variables in the public data file, that contains participants responses; b) the question frame that precedes each item; c) the items participants interacted with to produce responses; d) the response space the participants used to indicate their responses; e) the values used to code participants responses, and f) if any of the items were reverse coded before score generation.

The variable names are the names of the columns in the public data file that contains the responses of participants from the context questionnaire of TIMSS 2019 study. The variables used to generate the "Students Like Learning Science" scores are BSBS22A, BSBS22B, BSBS22C, BSBS22D, BSBS22E, BSBS22F, BSBS22G, BSBS22H, and BSBS22I. Each of these variables contains the responses to the respective items from the "Students Like Learning Science". For example, the variable BSBS22A stores participants responses to the item "I enjoy learning science". This item has four response categories: "Agree a lot", "Agree a little", "Disagree a little" and "Disagree a lot". Each of these categories was coded with the response values 1, 2, 3 and 4, thus a higher number indicates a higher degree of disagreement. The question frame that precedes the item is "How much do you agree with these statements about learning science".

Figure 10. instrument embedded codebook for "Students Like Learning Science"



Source: Chapter 16: Creating context questionnaire scales TIMSS 2019 (Yin & Fishbein, 2020)

In the present guidelines, we favour this type of codebook documentation, because with this information a secondary user has all the information needed to implement a scoring process. Thus, the recommended elements of information for codebooks of multi-item instruments are a) variable names; b) frame; c) items; d) response space; e) response values; and f) reverse flags. The recommend elements assures users have all the necessary information to interpret the collected responses from a study, generate scores, and produce results.

How to build a codebook

A study that collected responses to measure an SDG target, would benefit from the generation of data embedded codebooks and instrument embedded codebook, at the least. The first, assures users of the data file can access metadata to interpret what each value means in the shared data of the study. The second codebook, assures users of the study data files have enough information for many purposes, including the generation of scores for assessing SDG targets.

Data embedded codebooks can be generated using statistical software. Statistical software such as SAS, SPSS, STATA and R have commands to include metadata onto data tables and save this information into their data files. However, before building a codebook in a statistical package, most often researchers and analysts may create spreadsheets containing the basic information of the data file, for each variable. The basic elements contained in these spreadsheets are variable name, variable label, value labels including missing coded responses (Wu et al., 2016a, p. 65). For every variable contained in the data file generated for use, a row should be included to document its basic properties: name, label, type, values.

Figure 11. Spreadsheet codebook example of ICCS 2016 (selected fields)

ID	Variable	Label	Level	Range Minimum	Range Maximum	Value Scheme Detailed
9041	IDCNTRY	Participant Code	Nominal			
9262	IDSTUD	STUDENT ID	Nominal	10010101	94999999	
9261	IDSCHOOL	SCHOOL ID	Nominal	1001	9496	
9091	S_AGE	Student age	Ratio			
9183	S_GENDER	Student gender	Nominal			0: Boy; 1: Girl

Source: ICCS 2016 public data, see https://www.iea.nl/index.php/data-tools/repository/iccs, see ICCS2016MS_Codebook.xlsx

Instrument embedded codebooks are a friendlier form of documentation that can aid data inquiry. These codebooks serve the purpose of making it easier for users to find the name of a variable, once it is matched to the test or question survey that generate its responses. We recommend generating these codebooks, so users of the study data files are aware of items that belong to a scale, its reverse items, and what participants interacted with to generate responses. To generate these documents, word processors are needed, and a copy of the study instrument, so response values and variable names can be overlaid on top of the instrument in question. One limitation of Instrument embedded codebook should be noted. these latter documents are not designed to store information about study process variables such as students id, country codes, stratification variables, survey weights among other variables. These later process variables need to be documented in the succinct codebook, in a spreadsheet for example.

Very complete examples of these documents can be consulted in ICCS 2016 User guide (Köhler et al., 2018), TIMSS 2019 technical report (Martin et al., 2020) and PISA 2018 website[10].

[10] https://www.oecd.org/pisa/data/2018database/
https://webfs.oecd.org/pisa2018/PISA2018_CODEBOOK.xlsx

9. Producing scores

Calculation method

Proposing scores to assess SDG thematic indicator 4.4.2 using Large Scale Assessment data, requires the identification of available that can represent these indicators. Sandoval-Hernández et al. (2021) carried out a mapping exercise where SDG 4.4.2 indicator was mapped onto available measures from different large scale assessment studies including PISA, ICILS and PIAAC. The mapping exercise identified the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) as the most valuable source of information for SGD indicator 4.4.2. This study was chosen due to its conceptual framework (OECD, 2012), which showed the highest coverage of the topics relevant to this indicator. Additional reasons for the selection of PIAAC were that its target population covers the two groups mentioned in the indicator (youth and adults); as well as its potential to inform long-term monitoring.

Since the test design for PIAAC is based on a variant of matrix sampling (using different sets of items, multistage adaptive testing, and different assessment modes) where each respondent was administered a subset of items from the total item pool. The responses to the subset of test items are scaled using item response theory (IRT) methodology and combined with other background information (provided by the respondent) and model parameters to produce a set of 10 plausible values (PVs). These PVs can be used to produce group-level estimations of proficiency values (OECD, 2013).

According to the PIAAC Technical Report (OECD, 2013), the following steps can be followed to calculate an estimate T of the proficiency values Θ using PVs and to calculate an estimate of the variance of T:

1. Using the first vector of plausible values for each respondent, evaluate T as if the plausible values were the true values of Θ. Denote the result T_1.

2. In the same manner as in step 1 above, evaluate the sampling variance of T, or $\mathrm{Var}(\mathrm{T}_1)$, with respect to respondents' first vectors of plausible values. Denote the result Var_1.

3. Carry out steps 1 and 2 for the second through all 10 vectors of plausible values, thus obtaining T_v and Var_v for v=2,. . ., 10.

4. The best estimate of T obtainable from the plausible values is the average of the 10 values obtained from the different sets of plausible values:

$$\mathrm{T}_{\cdot} = \frac{\sum_{v} \mathrm{T}_{v}}{10} \tag{1}$$

5. An estimate of the variance of T is the sum of two components: an estimate of Var(T_{v}) obtained as in step 4 and the variance among the T_{v}s:

$$Var\mathrm{T}_{\cdot} = \frac{\sum_{n} Var_{v}}{10} + \left(1 + \frac{1}{10}\right)\frac{\sum_{v}(\mathrm{T}_{v} - \mathrm{T}_{\cdot})^{2}}{10 - 1} \tag{2}$$

The first component in $Var\mathrm{T}_{\cdot}$ reflects uncertainty due to sampling from the population; the second component reflects uncertainty because the respondents' proficiencies Θ are only indirectly observed.

Then, using the cut-off points established for the scale (see below), the proportion of students respondents reaching the corresponding standard is estimated within each country or region as a simple proportion (*P*).

$$P = \frac{X}{n} \tag{3}$$

Where X is the number of respondents that reach the standard in each country and n is the total number of respondents in the same country.

Definition of cut-off points (standards)

The performance of the participants in PIAAC-PSTRE is used to produce a proficiency scale (i.e., score) that ranges from 0 to 500. This scale is then divided into four proficiency levels (i.e., below 1, 1, 2 and 3) based on the knowledge and skills required to complete the tasks within those levels. Respondents at a particular level not only demonstrate knowledge and skills associated with that level but also the proficiencies required at lower levels. So, for example, respondents scoring at Level 2 are also proficient at Level 1.

To create the proficiency levels, an expert group in problem-solving in technology-rich environments met with psychometricians and test developers and reviewed data, looked at the tasks along the 500-point scale, and determined the requisite skills and knowledge to complete those tasks progressively increased along the scale. These proficiency levels of PSTRE are defined as shown in

Table 16.

Table 16. Description of the PSTRE proficiency levels.

Below Level 1 (0 to240 score points) Tasks are based on well-defined problems involving the use of only one function within a generic interface to meet one explicit criterion without any categorical, inferential reasoning or transforming of information. Few steps are required and no subgoal has to be generated.
Level 1 (241 to 290 score points) At this level, tasks typically require the use of widely available and familiar technology applications, such as email software or a Web browser. There is little or no navigation required to access the information or commands required to solve the problem. The problem may be solved regardless of one's awareness and use of specific tools and functions (e.g., a sort function). The task involves few steps and a minimal number of operators. At a cognitive level, the person can readily infer the goal from the task statement; problem resolution requires one to apply explicit criteria; there are few monitoring demands (e.g., the person does not have to check whether he or she has used the adequate procedure or made progress toward the solution). Identifying contents and operators can be done through simple match; only simple forms of reasoning, for example, assigning items to categories are required. There is no need to contrast or integrate information.
Level 2 (291 to 340 score points) At this level, tasks typically require the use of both generic and more specific technology applications. For instance, the person may have to make use of a novel online form. Some navigation across pages and applications is required to solve the problem. The use of tools (e.g., a sort function) can facilitate the resolution of the problem. The task may involve multiple steps and operators. In terms of cognitive processing, the problem goal may have to be defined by the person, though the criteria to be met are explicit. There are higher monitoring demands. Some unexpected outcomes or impasses may appear. The task may require evaluating the relevance of a set of items to discard distractors. Some integration and inferential reasoning may be needed.
Level 3 (341 to 500 score points) At this level, tasks typically require the use of both generic and more specific technology applications. Some navigation across pages and applications is required to solve the problem. The use of tools (e.g., a sort function) is required to make progress toward the solution. The task may involve multiple steps and operators. In terms of cognitive processing, the problem goal may have to be defined by the person, and the criteria to be met may or may not be explicit. There are typically high monitoring demands. Unexpected outcomes and impasses are likely to occur. The task may require evaluating the relevance and the reliability of information in order to discard distractors. Integration and inferential reasoning may be needed to a large extent.

Source: PIAAC Technical Report (OECD, 2013)

By comparing the definition of SDG Indicator 4.4.2 and the description of the problem-solving in technology-rich environments proficiency levels, we identified Level 2 as the threshold or cut-off point to estimate the proportion of respondents reaching the indicator within each country. At Level 2, tasks typically require the use of both generic and more specific technology applications.

At the threshold, respondents typically require the use of both generic and specific technology applications. Adults at this level are typically able to use software they have never seen before to solve problems, even when unexpected impasses/outcomes occur. For example, they are likely able to:

- Figure out how to send an email message to a number of contacts using an unfamiliar bulk email function;
- Use a sorting tool to make it easier to locate sales numbers for a specific product in a company spreadsheet;
- Conduct a web search to find out how to solve a problem with other software, such as how to view a column that won't display properly in a spreadsheet; and
- Find an email message or file that has been "lost" somewhere on a computer hard drive.

It is very important to notice that the information to produce the scores to monitor SDG 4.4.2 cannot be replicated without having access to the items used in PIAAC to measure the PSTRE dimension. At the moment these items are not publicly available and access and permission to use them should be negotiated with the PIAAC team at the OECD.

10. Using the results of the national assessment

We have compiled in this document a set of guidelines for countries to implement a national assessment that allows them to produce information to measure and monitor SDG 4.4.2. This includes all the major phases that national and international assessments incorporate, such as deciding who will carry out the assessment, the objectives of it, the definition of the population to be assessed, the development of the assessment framework, logistic considerations for the data collection (e.g. development of manuals), the sampling, weighting and variance estimation procedures, data preparation and management (e.g. scoring) and the reporting of the results of the assessment.

We have also provided detailed instructions on how to conduct all these phases of the assessment and have provided examples and exercises to facilitate the tasks for implementation agencies. We have focused on state-of-the-art procedures that need to be followed in order to ensure that the data produced by the assessment exercise are of high quality and address the concerns of policymakers, decision-makers, and other potential users of the information.

These Guidelines are intended primarily for the teams within the designated implementation agencies who are responsible for conducting a national assessment exercise.

As readers make their way through these Guidelines, it will become evident that the successful implementation of a national assessment exercise is a complex task that requires considerable knowledge, skill, and resources. Good quality implementation of these Guidelines will increase the confidence of policymakers and other stakeholders in the validity of the information produced. It also can increase the likelihood that the results of the national assessment will be used to develop educational plans and programmes.

References

ACER. (2018). *Replicates* (9.2). ACER.

Anderson, P., & Morgan, G. (2008a). *Developing Tests and Questionnaires for a National Assessment of Educational Achievement*. World Bank.

Anderson, P., & Morgan, G. (2008b). The test administrator. In P. Anderson & G. Morgan (Eds.), *Developing Tests and Questionnaires for a National Assessment of Educational Achievement*. World Bank.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2014). R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics. *Technology Innovations in Statistics Education*, *8*(1), 20.

Brese, F., & Cockle, M. (2017). Data Management Procedures. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of Large-Scale Education Assessments* (pp. 253–275). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118762462.ch10

Brese, F., Jung, M., Mirazchiyski, P., Schulz, W., & Zuehlke, O. (2011). *ICCS 2009 User Guide for the International Database*. International Association for the Evaluation of Educational Achievement (IEA).

Brick, M. J., Morganstein, D., & Valliant, R. (2000). *WesVar* (5.1). Westat.

Caro, D., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software*, *81*(7), 1–44.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427–444.

Desjardings, C. D., & Bulut, O. (2018). *Handbook of Educational Measurement and Psychometrics Using R*. CRC Press, Taylor & Francis Group.

Dumais, J., & Gough, H. (2012). Computing estimates and their sampling errors from complex samples. In V. Greaney & T. Kellaghan (Eds.), *Implementing a National Assessment of Educational Achievement*. World Bank.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 Assessment Framework*. Springer Open. 10.1007/978-3-030-19389-8

Gebhardt, E., & Berezner, A. (2017). Database Production for Large-Scale Educational Assessments. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of Large-Scale Education Assessments* (pp. 411–423). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118762462.ch16

Gonzalez, E. J. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments*, *5*, 115–134.

Greaney, V., & Kellaghan, T. (2008). *Assessing National Achievement Levels in Education*. World Bank.

Greaney, V., & Kellaghan, T. (2012). *Implementing a National Assessment of Educational Achievement*. World Bank. https://doi.org/10.1596/978-0-8213-8589-0

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Heeringa, S. G., West, B., & Berglund, P. A. (2009). *Applied Survey Data Analysis*. Taylor & Francis Group.

Howie, S., & Acana, S. (2012). Preparation for adminsitration in schools. In V. Greaney & T. Kelly (Eds.), *Implementing a National Assessment of Educational Achievement*. World Bank.

IBM. (2015). *IBM SPSS statistics for windows* (No. 23). IBM Corp.

IEA. (2017). *IEA Studies in Ten Steps*. Youtube.

IEA. (2019). *Help Manual for the IEA IDB Analyzer (Version 4.0)*.

Kellaghan, T., & Greaney, V. (2001). *Using Assessment to Improve the Quality of Education*. International Institute for Educational Planning.

Kellaghan, T., & Greaney, V. (2004). *Assessing Student Learning in Africa*. World Bank.

Kellaghan, T., Greaney, V., & Murray, T. S. (2009). *Using the Results of a National Assessment of Educational Achievement*. World Bank.

Köhler, H., Weber, S., Brese, F., Schulz, W., & Carstens, R. (2018). *ICCS 2016 User Guide for the International Database* (H. Köhler, S. Weber, F. Brese, W. Schulz, & R. Carstens, Eds.). International Association for the Evaluation of Educational Achievement (IEA).

Laanpere, M. (2019). *Recommendations on assessment tools for monitoring digital literacy within unesco's digital literacy global framework*. UNESCO Institute for Statistics. http://uis.unesco.org/sites/default/files/documents/ip56-recommendations-assessment-tools-digital-literacy-2019-en.pdf

Law, N., Woo, D., de la Torre, J., & Wong, G. (2018). *A global framework of reference on digital literacy skills for indicator 4.4. 2* (UIS/2018/ICT/IP/51). UNESCO Institute for Statistics. http://uis.unesco.org/sites/default/files/documents/ip51-global-framework-reference-digital-literacy-skills-2018-en.pdf

Lee, H. (1995). Outliers in business surveys. In B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, & P. Kott (Eds.), *Business Survey Methods* (pp. 503–526). John Wiley & Sons.

Lietz, P., Cresswell, J., Rust, K., & Adams, R. (2017). *Implementation of large-scale education assessments*. Wiley.

Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. TIMSS & PIRLS International Study Center and International Association for the Evaluation of Educational Achievement (IEA).

Masters, G. N. (2016). Partial Credit Model. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory. Volume One. Models* (pp. 109–126). CRC Press.

Miranda, D., & Castillo, J. C. (2018). *Measurement Model and Invariance Testing of Scales Measuring Egalitarian Values in ICCS 2009* (A. Sandoval-Hernández, M. M. Isac, & D. Miranda, Eds.; Vol. 4, pp. 19–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-78692-6_3

Mirazchiyski, P. (2021). *R Analyzer for Large-Scale Assessments (RALSA)* (v0.90.2). INERI.

Mohadjer, L., Krenzke, T., Van de Kerckhov, W., & Li, L. (2013). Sampling Design. In *Technical Report of the Survey of Adult Skills (PIAAC)*. OECD Publishing. https://www.oecd.org/skills/piaac/publications/PIAAC_Technical_Report_3rd_Edition_2019_Section4_Chapters14-16.pdf

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8.2). Muthén & Muthén.

Napier, J. L., Thorisdottir, H., & Jost, J. T. (2010). The joy of sexism? A multinational investigation of hostile and benevolent justifications for gender inequality and their relations to subjective well-being. *Sex Roles*, *62*(7–8), 405–419. https://doi.org/10.1007/s11199-009-9712-7

OECD. (2012). *Literacy, Numeracy and Problem Solving in Technology-Rich Environments. Framework for the OECD Survey of Adult Skills*. OECD Publishing. https://doi.org/10.1787/9789264128859-en

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing.

OECD. (2021). *PISA 2018 Technical Report*. OECD Publishing.

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education*, *9*(1), 1–17. https://doi.org/10.1186/s40536-020-00092-z

R Development Core Team. (2011). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Rocher, T., & Hastedt, D. (2020). International large-scale assessments in education: A brief guide. *IEA Compass: Briefs in Education*, *10*.

Rust, K. F. (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 Background.Technical Issues, and Methods of Data Analysis*. CRC Press.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2014). *Handbook of International Large-Scale Assessment*. CRC Press.

Sandoval-Hernández, A., & Carrasco, D. (2020). *A Measurement Strategy for SDG Thematic Indicators 4.7.4 and 4.7.5 Using International Large Scale Assessments in Education*. UNESCO Institute for statistics.

Sandoval-Hernandez, A., Isac, M. M., Carrasco, D., & Miranda, D. (2021). *Guidelines for Data Collection to Measure SDG 4.7.4 and 4.7.5.* UNESCO Institute for Statistics.

Sandoval-Hernández, A., Isac, M. M., & Miranda, D. (2019). *Proposal of a Measurement Strategy for SDG Global Indicator 4.7.1 and Thematic Indicators 4.7.4 and 4.7.5 using International Large-Scale Assessments in Education*. UNESCO Institute for statistics.

Sandoval-Hernandez, A., Osorio-Saez, E., & Eryolmaz, N. (2021). *Measurement Strategy for SDG Global Indicator 4.4.2 Using International Large-Scale Assessments*. UNESCO Institute for Statistics.

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer.

SAS. (2012). *SAS System for Windows (Version 9.4)*. SAS Institute.

Schulz, W., Ainley, J., & Fraillon, J. (2011). *ICCS 2009 technical report*. International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2018a). *ICCS 2016 Technical Report*. The International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2018b). *ICCS 2016 technical report*. International Association for the Evaluation of Educational Achievement.

Shiel, G., & Cartwright, F. (2015). *Analyzing Data from a National Assessment of Educational Achievement*. World Bank.

Wilson, M., & Draney, K. (2002). A Technique for Setting Standards and Maintaining Them over Time. *Measurement and Multivariate Analysis*, 325–332. https://doi.org/10.1007/978-4-431-65955-6_35

Wu, M., Tam, H. P., & Jen, T.-H. (2016a). *Educational Measurement for Applied Researchers*. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5

Wu, M., Tam, H. P., & Jen, T.-H. (2016b). Partial Credit Model. In *Educational Measurement for Applied Researchers* (pp. 159–185). Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5_9

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 Context Questionnaire Scales. In M. O. Martin, M. Von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Appendix I. Sample items from PIAAC instruments that were used to evaluate Problem-Solving in Technology-Rich Environments (PSTRE)

The examples of the PSTRE items presented below are taken from the PIAAC Framework (OECD, 2012, pp. 53–55). Please note that these items are administered in electronic format, and these examples correspond to the print layout.

Sample item 1

In this item, respondents must access and evaluate information in the context of a simulated job search. The instructions, located on the left side of the screen, require respondents to identify and then bookmark one or more sites that do not require users to register or pay a fee.

Figure 12. Screenshot 1 of sample item 1.

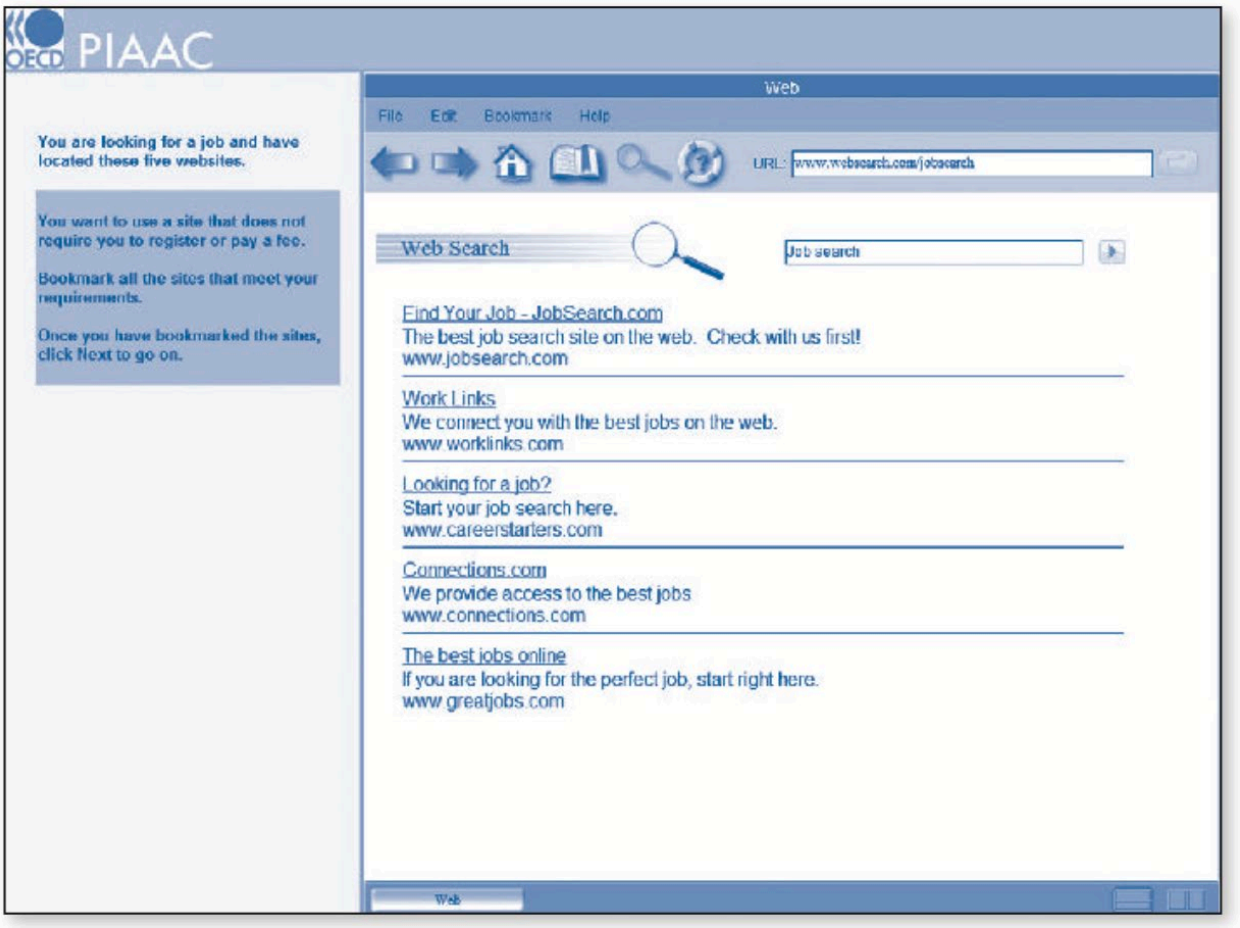
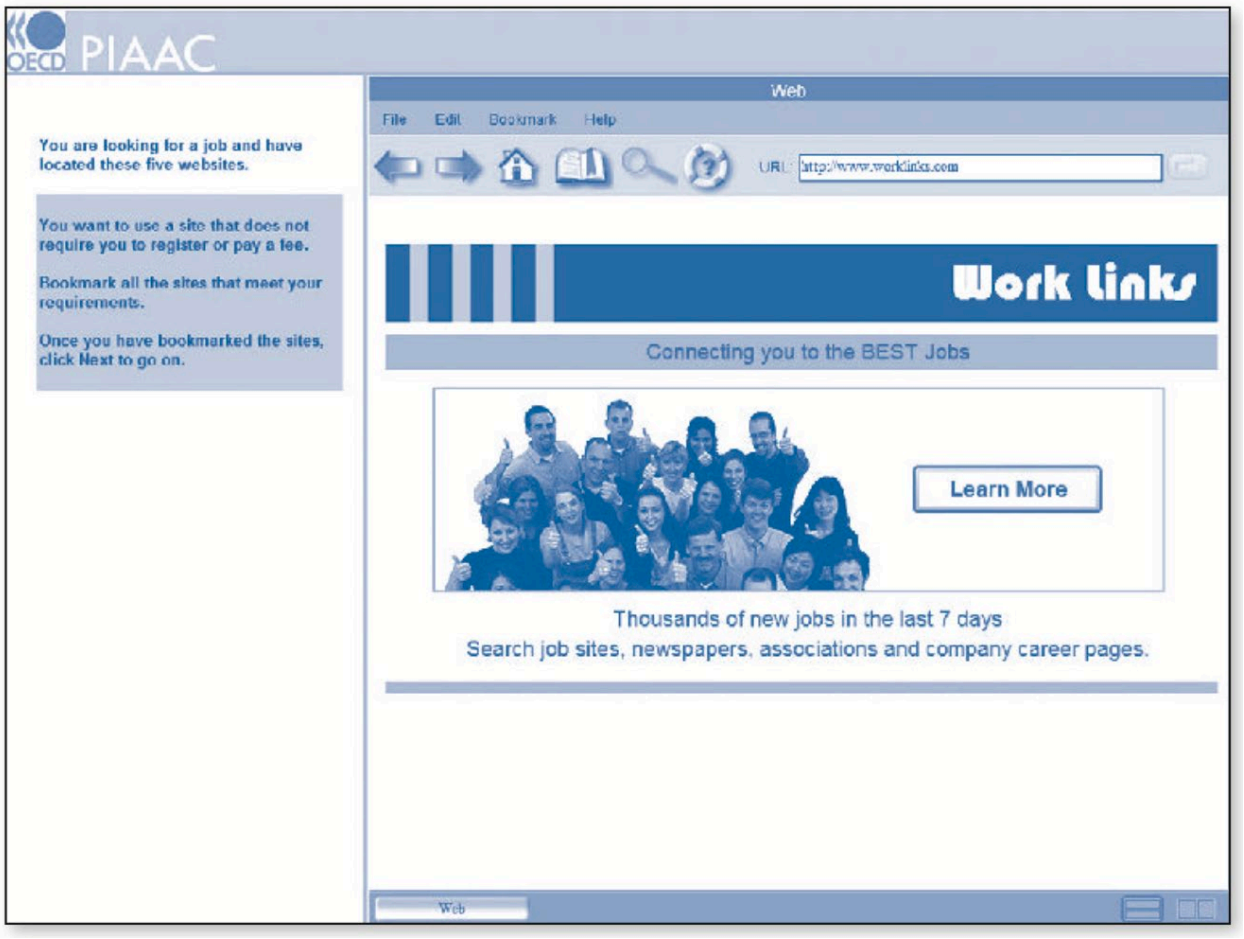


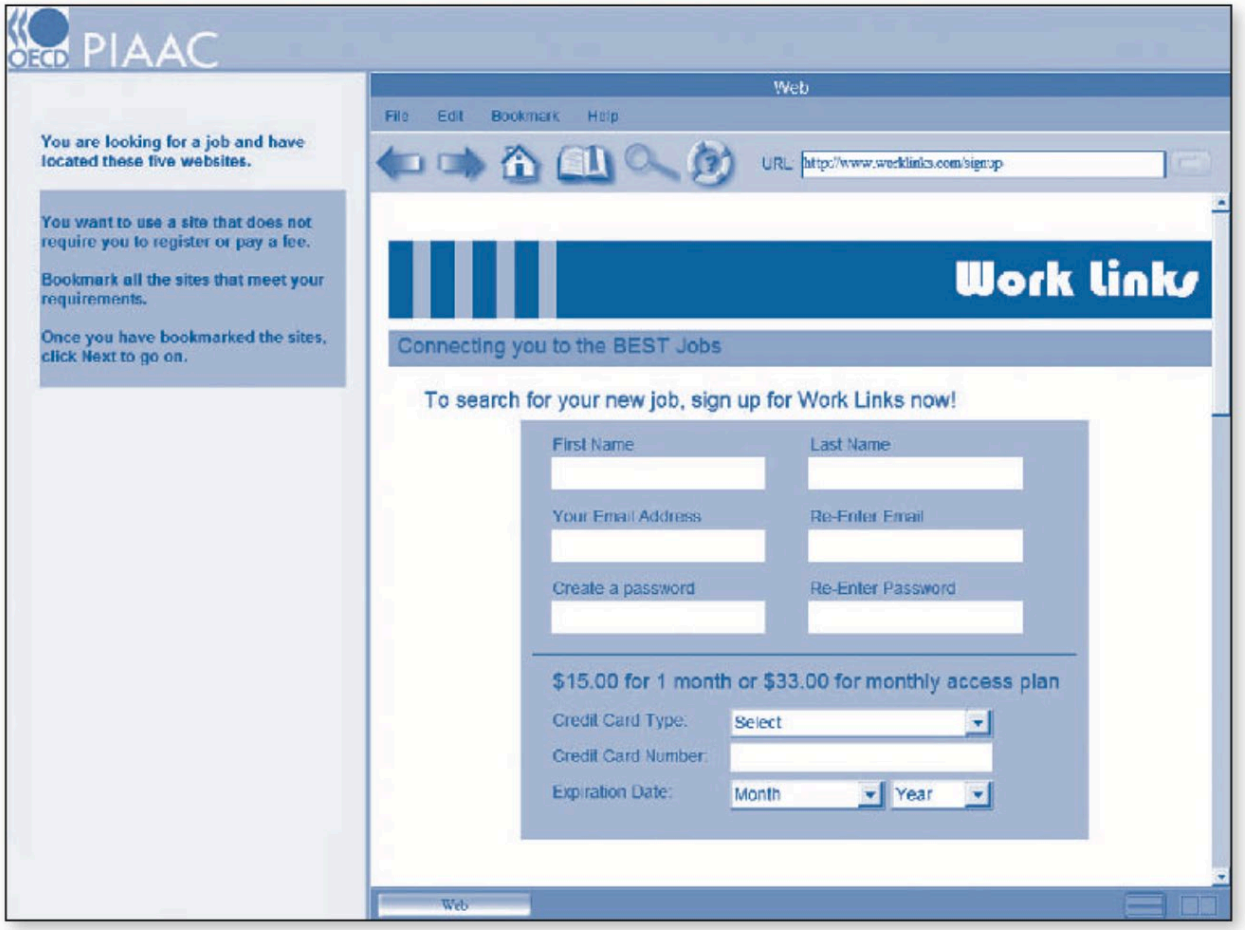
Figure 13. Screenshot 2 of sample item 1.



As can be seen, this item requires that respondents work within a simulated web environment that includes tools and functionality similar to those found in real-life applications. users are able to:

- Click on links on both the results page and associated web pages;
- Navigate using the back and forward arrows or the home icon; and
- Bookmark web pages and view or change those bookmarks.

Figure 14. Screenshot 3 of sample item 1.



In order to perform this task correctly, respondents may have to search through several pages on a website. One of the features of PIAAC is that the process and paths by which a respondent responds to the tasks are captured. For example, one of the websites, presented below, does not meet the criteria of not requiring registration or the payment of a fee, but the relevant information is not on the opening page. If a respondent bookmarks this site without clicking on the "Learn more" link to view the relevant information (see the website on the following page), this response may be interpreted in a different way than if the relevant page had been viewed. The breadth of information, combined with frameworks that specify behaviours of interest, allow PIAAC to learn more about what adults know and can do relative to the construct of problem-solving.

The relevant information is located on the form that indicates that users must sign up (register) and pay a fee.

Sample item 2

In this item, respondents select a set of files to download onto a portable music player. The files must meet specified criteria in terms of genre (jazz and rock) and not exceed the capacity of the device (maximum of 20 MB).

The software includes an automatic summing functionality ("total Size Selected") that facilitates the task by updating the total file size as files are selected or de-selected. Respondents must monitor progress as they select files, checking against the specified criteria to know when they have satisfied the constraints presented in the problem.

It is also possible to sort the spreadsheet by file size and/or genre, a strategy that can improve task efficiency. The connection between the use of resources in a technology-rich environment and resulting efficiencies for solving problems is emphasised in the framework and therefore included across items in the assessment.

Figure 15. Screenshot 1 of sample item 2.