United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
for
STATISTICS

UIS/OLOM/CLOM/2014
Montreal, April 2015

Survey 2015
Observatory of Learning Outcomes

INSTRUCTION MANUAL FOR COMPLETING THE CATALOGUE OF LEARNING ASSESSMENTS

CONTENTS

Technical notes for the completion of the template

1) Notes on coding missing data

The correct use of codes for missing information is essential to ensure the data's integrity. The reasons for which data is not available need to be identified and differentiated in statistical analyses and reports.

Do not leave any cell reserved for writing in a response blank. Cells with no data or invalid data values must be assigned one of the following five codes:

i) Category not applicable = **a**

Assign code **"a"** to cells or categories that do not apply to a given assessment or public examination. This implies that data for these categories or cells do not exist.

ii) Quantity nil = **n**

Assign code **"n"** to cells for which the data value (or quantity) is nil or negligible. Therefore, to indicate that a value is nil, use the code **"n"**, as opposed to the value 0 (zero).

iii) Data included in another category = **x**

Assign code "**x**" to cells where data cannot be disaggregated. **Please specify** in the metadata space where and if the data has been included in the template.

iv) Data not available = **m**

Assign code **"m"** to cells or certain classification categories where data is missing or not available, and for which data is not included in any other cells of the template (even though these data could, in principle, be collected).

iv) Data is confidential = **c**

Assign code "c" to confidential data that the country wants to refrain from sharing publically.

2) Note on provisional or estimated data = *

Please indicate any provisional or estimated figures by preceding the data value with an asterisk (*).

3) Note on partial data = p

Please indicate any partial figure by preceding the data value with a code (p).

1. INTRODUCTION

The UNESCO Institute for Statistics (UIS) has launched a new global initiative to map the different assessments countries use to monitor student achievement. Global education consultations have highlighted that in addition to ensuring equitable access to quality education, learning has a place on the post-2015 development agenda. Within the area of learning, competencies and skills for several domains have been identified as important for children and youth to acquire in order for them to lead productive lives and face future challenges. National assessments, both high- and low-stakes, are common in most countries. Documenting, analyzing, and disseminating the vast amount of knowledge these assessments yield is a core feature of the UIS Catalogue on learning outcomes measurement. The UIS will make the contents of the Catalogue available through its international database. Additional information is available, on the UIS website.

The Catalogue will be a global inventory of all national and international large-scale assessments and public national examinations that countries use to measure and monitor learning outcomes in primary and lower secondary schooling (ISCED 2011, levels 1 and 2). The Catalogue will be produced based on information submitted through a data entry template (questionnaire) that is designed to capture information on the characteristics of available student assessments. This template compiles data on the most recently administered nationally representative assessments between 2000 and 2014. Assessments administered before 2000 should be excluded from this Catalogue.

Outputs. The inventory of information on student assessments generated through the Catalogue will be added to the UIS international data base, and hence will be publically available. The information from the Catalogue will be combined with data already available in the UIS international database to produce brief country summaries that portray country efforts to measure and monitor student achievement. It will include the list of assessments - national or cross-national, high-or low stakes-, the domains they cover, and the extent to which school children are learning in these different domains. The country summary will provide key descriptive information about these assessments, and differences and similarities between the assessment attributes without comparing the quality of assessments within or across countries.

Expected outcomes. The country summaries can be used for informed dialogues about the use of assessment among various stakeholders. Governments and local institutions may use the summaries as relevant background information to review the characteristics of their assessments against specific objectives. The development community, such as bilateral and multi-lateral funding organizations, may consider the country summaries as part of a toolkit generally used to evaluate proposals in support of assessing learning outcomes. The UIS will take an active role in fostering the use of the country summaries and the information in the international database; for example, to inform self-initiated country reviews of scope and coverage of student assessments.

Benefits to countries. Preliminary estimates by the UIS indicate that more than 150 countries currently measure learning levels through national assessments and examinations or participate in international assessment initiatives (including regional or cross-national assessments). So far, there has been no global effort to gather systematic information about the assessment choices that countries make and how these relate to the learning outcomes agenda. The Catalogue will fill this gap by centrally collating the current wealth of assessment data and producing country summaries. Additionally the UIS will engage in further research that may be helpful to country decision makers such as tracking and documenting emerging assessment trends and patterns across countries, without appraising their quality or comparing results across countries.

Governments may be motivated to review the characteristics of student assessments for various reasons including

(i) the fact that high-quality assessments tend to be expensive on a per capita basis; therefore, it may be desirable to examine the utility and value added of each assessment;
(ii) Public confidence in testing may be influenced by the public perception of the fairness of testing, which in turn may be associated with the quality of the test; and
(iii) Countries may want to align their assessments with international standards. The Catalogue outputs can thus be a useful resource to support such revisions.

As part of the UIS dissemination strategy for the Catalogue country summary and research reports, the UIS plans to offer countries the opportunities to join communities of practice, in which governments may share their experiences on monitoring

learning outcomes. The expectation is a clear focus on national assessments. The planned communities of practice may also serve as forum for discussions around country systems and emerging global standards.

The Catalogue will not compare student performance. Student performance results are not necessarily comparable, especially across countries' national assessments, since these are not designed to be comparable. However, comparability should not be considered a pre-requisite for analyses of assessment results. Non-comparable data are valuable in monitoring learning outcomes—nationally and internationally.

Irrespective of comparability, the documentation of student assessments drawn from the existing pool of databases has proven to be influential in stirring interest in the reasons for differences across countries. It is important to point out that there could be several underlying factors for these differences that may not be immediately apparent.

While student performance results are not always comparable, their characteristics are which include among others: grades, approaches, content domains, languages, and how results are reported and used for policy-making. Descriptive analyses of assessment attributes reveal patterns of decisions that governments make to monitor learning outcomes. To inform global patterns and trends in assessment, the Catalogue will compare and contrast countries based on the descriptive attributes of their assessments.

This **Instruction Manual** has been prepared to assist data providers when completing the Catalogue of Learning Outcomes Measurement. For additional information on how to complete and submit the Catalogue please contact:
Mr. Georges Boade by e-mail at g.boade@unesco.org; or by telephone at 1-514-343-7845.

2. COVERAGE OF THE QUESTIONNAIRE

Types of assessments to be included in this survey:

This survey focuses on system-level educational assessments which primarily target students enroled in formal education programmes. Only data from large-scale national assessments, including public/national (exit or end-point) examinations administered in ISCED levels 1 and 2, in general or technical/vocational education programmes, will be collected in this template. For more details regarding ISCED levels and grades in your country, consult ISCED 97 mappings and the ISCED 2011 framework. The international assessments that countries participate in should be listed in Section 2.3, but the data will be collected directly from the source agencies; that is, data from international assessments will not be recorded in this template by country officials. Examples of international assessments include LLECE, TIMSS, PIRLS, PASEC and SACMEQ.

Types of assessments to exclude from this survey:

Three categories of assessments are not included in this survey:

1. School-based assessments and assessments organized by decentralised educational institutions which are mainly for selection purposes from one grade to another or one programme to another. These include:

    a. tests generally organized at the end of primary school with a purpose to select well-performing students into a first-year secondary general or technical/vocational programmes;

    b. continues or sequential evaluations organised by teachers, head of school institutions or the national autority in charge of the education

    c. tests administered to students for selection to professional or training schools; and

    d. assessments organised by professional or training schools

2. Tests which do not specifically target students enroled in education programmes and which are administered to all candidates who apply for a professional certification, generally without prior enrollment to an equivalent official general or vocational educational programme. This type of test is decentralised and can be organised by unions or professional associations. Driving license exams or exams for membership to a professional association are few examples.

3. Household-based assessments of learning outcomes, even if they target children enroled in ISCED level 1 or 2 programmes.

The template is designed to accommodate different country scenarios. For example:

- ✓ Section 2 has space for 22 assessments. If the standard form is too small for a specific country, please contact us to create more space.
- ✓ Section 7: 7.2.3 accommodates data for ten subjects/constructs covered in a single assessment. If an assessment covers more than 10 mandatory subjects/constructs, contact us to add more columns.

3. HOW TO COMPLETE THE QUESTIONNAIRE

Respondents are requested to carefully read the following instructions, definitions and illustrations to accurately complete the UIS Catalogue of learning outcomes template. This manual covers all the definitions of key variables used in the template and explains the kind of data requested in each item or table.

The questionnaire is comprised of eight sections. Each section ends with an item called "Metadata". This space is included for respondents to provide any specific explanations or clarifications they have regarding this given section. It will also help with interpreting the data requested.

If the same information and data are required in more than one question; please ensure that your entries are consistent across questions. In cases where there is an unavoidable discrepancy, please explain this clearly in the Metadata field at the end of each section.

Please do not leave any cell where a written answer or response option is expected blank. The correct use of missing data **codes** is essential to ensure the integrity of the data. Reasons why there are no data in a particular instance need to be distinguished in UIS statistical analyses and reports.

The UIS encourages all countries to make their own estimations of missing or incomplete data. To signal that a cell contains estimated values, please put an asterisk (*) in front of the number, i.e. *68794. Do not leave a space between the symbol and the figure. If, despite these efforts, some data are not available or are incomplete, please explain in a separate note or in the metadata field notes at the end of each section.

In addition to the specific explanations or clarifications to be included in the Metadata sections, please use this space to cite any reference material, manual, publication or website that you have referred to, which may help in understanding the data of your country. The UIS encourages you to send those documents along with the completed template for a better understanding of the data.

Refer to the **Glossary** at the end of this manual for definitions and explanations of the data to be reported in the template.
Also, note that the catalogue survey seeks to collect information about assessments and public examinations administered between **2000 and 2014**. Please always indicate the year the data has been reported.

4. INSTRUCTIONS FOR EACH QUESTION

SECTION 2: List of Assessments

Report each assessment only once, even if it covers more than one grade or subject. Take distinct testing under the same name as one assessment, independent of how many distinct grades, subjects and languages it includes. Often the assessment will test more than one subject or construct per grade. As long as the same name is assigned to testing in multiple grades, these versions should be considered as being part of the same assessment. Then, if an assessment is administered in a same year to more than one grade, list it once only- this complementary information will be captured in the next sections.

This survey is about large scale assessments administered to students of primary schools and general/technical lower secondary education in your country. They cover the whole country and target all or sample of students that share one or more attributes like grade in which they are enroled, or age.

 *Please refer to previous pages of this manual for the* ***types of assessments to consider for the purpose of this survey****, and to the glossary for the definition of* ***low-stake*** *and* ***high-stake assessments,*** *before completing this section.*

 **2.1 List of large-scale national assessments**

This table seeks general information on <u>all</u> large-scale national assessments.
Please enter the following information for each assessment, in a chronological order of administration, starting with the most recently administered assessment:

i) the acronym in its **<u>original language</u>** if it exists (acronyms are never to be translated). Please assign a code "a" if the acronym does not exist.
ii) the complete name of the assessment:
   - in front of the letter N, in its original language.
   - in front of the letter E, its translation in English if available, when the original language is not English.
iii) the education programme(s) the assessment covers. Select from the list provided:
   - primary education only (ISCED 1)
   - lower secondary education only (ISCED 2)
   - primary and lower secondary education (ISCED 1 & 2)
   - lower and upper secondary (ISCED 2 & 3)
   - primary, lower and upper secondary education (ISCED 1 , 2 & 3)

iv) the name of the institutions responsible for the administration of the assessment.
v) the stake of each assessment listed, whether low or high for each of the following populations: students, teachers and schools. Note that there are some cases where the stake of the assessment is informal. If possible, include explanations in the metadata section.
vi) the latest year *(yyyy)* the assessment was administered.
vii) in case your country has a federal administrative structure, without any national assessment that covers all states, provinces or regions (depending on the terminology used in your country), please select Yes in the red cell under this table. That means table 2.1 will be empty as there is no large-scale national assessment.

Include any observations/clarifications that should be considered together with the data provided to relate a full and accurate picture in Observations about section 2.1.

2.2 List of national public examinations

This table seeks general information on national public examinations. The same instructions as for 2.1 apply; therefore, please refer to 2.1 for more detailed instructions.

In case your country has a federal administrative structure, without any national public examination that covers all the states, provinces or regions (depending on the terminology used in your country), please select Yes in the red cell under the table. That means table 2.2 will be empty as there is no national public examination.

Include any observations/clarifications that should be considered together with the data provided to relate a full and accurate picture in metadata about sub-section 2.2.

2.3 List of international assessments

This table seeks general information on the international assessments that have been administered in your country. This category includes global assessments or any assessment that is administered in two or more countries.

In case your country has a federal administrative structure and only some states, provinces or regions (depending on the terminology used in your country) have participated to the international assessments, please select Yes in the red cell under the table. That means table 2.3 will be empty as there is no international assessment that has been administered nationally, but only in a selected number of states, provinces or regions.

Include any observations/clarifications that should be considered together with the data provided to relate a full and accurate picture in Observations about section 2.3

Please note that country focal points are responsible for filling sections 3 and beyond with regards to the national assessments and public examinations in their country. Data about international assessments will be collected directly from the institution responsible for developing this assessment.

2.4 Existence of large-scale assessment in Early Childhood Education (ISCED level 0)

This sub-section aims to highlight the pre-primary or other early childhood education large-scale assessments that exist in your country. If you select **Yes**, indicating that there are large-scale assessments of learning outcomes in pre-primary programmes in your country, provide the title of each assessment, the age of children it targets, its coverage (national or sub-national) and the institutions responsible for its administration.

Include any observations/clarifications that could complement the data provided in Observations about section 2.4.

2.5 Existence of households-based assessments of learning outcomes

This sub-section aims to highlight the large-scale household-based assessment of learning outcomes that exist in your country.

Household-based assessments of learning outcomes are generally conducted by interviewers that visit households in order to collect background information on the household, and to assess the cognitive skills of one or more members of the household.

Therefore, do not include here school-based assessments where a child brings a questionnaire home to their parents. In that case, this questionnaire is part of the instruments related to the assessments listed in sub-sections 2.1 to 2.3.

Household-based assessments of learning outcomes do not only target students enroled in formal programmes; they generally include out-of-school children and may apply other selection criteria, such as age of children to be assessed.

From section 3 onwards, you are asked to provide information that will help characterise each national assessment and public examination listed in Section 2 (sub-sections 2.1 and 2.2 only).

Data on international assessments (listed in subsection 2.3) will be completed by the institution that is responsible for its administration.

Before you start completing the data for these sections, please make as many copies of this template as the number of assessments listed in sections 2.1 and 2.2. To easily identify the templates, ensure that each copy indicates the country name, followed by the assessment acronym if it exists (otherwise, you can choose any name that can easily differentiate the assessment). For example: Sri Lanka-GCE-O-L or France-CAP.

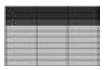
Then, start completing these sections for each assessment

SECTION 3: Scope, purposes, funding and stakeholders

This section provides the information that will characterise each assessment. It includes data on the scope of the assessment, its purposes, assessment funding sources and the roles of each stakeholder.

Green cells are preloaded with a list of options to select from; hence, there is no need to type in your responses.
In some cases where it is required to specify a category not listed <*other, please specify*>, a **yellow cell** indicates the place to type that category.
If there are observations that should be considered to understand these data, please write them in the metadata space available at the end of this section.

To complete this section, you may need to use or explore the latest official documentation and archives such as the assessment frameworks, reports and databases.



3.1. General information about the target assessment or public examination

Please write the full name of the assessment or examination in English and in its national or official language if not English, as well as the acronym, exactly as in section 2. If the acronym does not exist, please assign a code "a" to item 3.1.3.

*Please refer to the glossary for the definitions of **high-stakes and low-stakes assessments.***

3.2. Scope of the assessment

A given assessment may be designed to cover several categories within each variable. For example, an assessment may be designed to cover several grades or children of a certain age range. Please select or provide all categories that apply per variable.

3.2.1-3.2.2: For each of the variables listed below, tick the category or categories that apply to the assessment:
3.2.1. education programme(s) /ISCED levels
3.2.2. grade (s)
3.2.3. age groups
3.2.4. target curriculum. If other, please specify in the **yellow cell**, and include necessary explanations in the metadata space at the end of the section.
3.2.5. types of school institutions. If other, please specify that other type of school in the **yellow cell**, and provide a definition in the space provided for metadata.

3.2.6. Please select the programme orientation (technical/vocational, general or both). Assessment usually target either a general or vocational/technical education programme; however, in the event that the assessment covers both, please write any necessary explanation in the metadata notes at the end of the section.

3.2.7. Please select the format in which the assessment (or the test) is delivered (written, oral, practical or a combination). For assessments that require students to only take written tests, select Yes for "written"; on the other hand, when the proficiency of the student is determined based on oral tests only, select Yes for "oral". If the assessment is based on a practical activity, please select Yes for "practical". If the test includes additional formats, please specify them in the **yellow cell** and include necessary explanations in the metadata notes at the end of the section. If the assessment (or the test) includes more than one test format (written and oral; or written and practical; or the three formats at the same time), please provide clarification in the space provided for metadata to indicate which part of the assessment is delivered in what format, and any other additional information.

*Please refer to the glossary for the definitions of **grade, ISCED levels and construct.***

3.3. Purposes of the assessment

Please select the purpose(s) intended by the assessment. Select all that apply and in the **yellow cell**, add any additional purpose intended by the assessment, not included in the pre-defined list.

Please give any additional information for each relevant purpose in the space provided for comments.

3.4. Funding sources

Please select the funding source(s) that apply to this assessment. Select all that apply and in the **yellow cell**, add any additional funding source, not included in the predefined list. There may be several sources of funding involving a number of stakeholders. Please provide the disaggregated information if available in the space provided for comments, or at the end of the section. An example of disaggregation could be to show the proportion of overall funding by funder.

Note: For the purpose of this questionnaire, foreign organizations are those created and mainly funded by citizens of another country, while the local organizations refer to those created, funded and managed by citizens and residents of your country. International organizations refer to bilateral or multilateral funded and managed organizations. For this reason, national offices of multilateral organizations such as the World Bank, Food and Agriculture Organization, UNICEF or UNESCO are considered international organizations even if funding is mainly coming from the country where they are established.

If possible, please indicate the name of the internations organizations involved in the funding of the assessment (or the test). Please give any additional information for each relevant funding source in the space provided for comments.

3.5. Stakeholders and their roles

An indicative list of potential stakeholders is provided in rows, and the possible stakeholder roles in columns. Please select the applicable stakeholder and role combination. It is possible for one stakeholder to have multiple roles. Additionally, a given role may be shared among several stakeholders. Please select all that apply and add any other stakeholder implicated in the assessment in the **yellow cell** of the last row, and any other role not accounted for in the **yellow cell** of the last column.

 In countries with a **federal administrative structure**, the situations may vary significantly from one country to another and one region within the country to another. In this case, please provide in the metadata space all information that is necessary to understand and document the case for data users.

SECTION 4: Test design and administration

4.1. Test design

4.1.1. A single booklet is administered to all students when the same content is tested. However, there are cases where students participating in a given assessment are administered different content; this may happen when there are too many items to be administered. Given time constraints and the added burden to test takers, only a subset of the item bank is administered to each student. Students therefore receive different test booklets that are comprised of a selection of items; the booklets are assembled by performing a method known as complex matrix sampling and there is a subset of common items in each booklet.

In some cases, test booklets are developed so that a different group of students are tested on a subset of the overall learning objectives. In this case, test booklets do not have a common subset of items, and are based on different learning objectives.

If another selection method was applied for this assessment or public examination, please specific in the **yellow cell** of the last column of this table.

Please select the assessment's test selection method. Please refer to the glossary for the definition of complex **matrix sampling.**

4.1.2. In this table, you are asked to provide the delivery model(s) and the mode(s) of data collection that have been used to administer this assessment. For each relevant delivery model, please select the appropriate mode(s) of data collection:

i. a face-to face mode of data collection is when students take the test under the presence of supervisors (assessment staff).
ii. an online mode of data collection is when students take the test via the internet without a need to be supervised by assessment staff. Note that the allowed time to complete the assessment may vary.
iii. a mail mode of data collection is when students receive the test via postal mail, and complete it without a need to be supervised by assessment staff.
iv. if a delivery model or a mode of data collection is not listed, please write it in the **yellow cell** of the last row or column respectively and give additional information if necessary in the metadata space at the end of the section.

Please refer to the glossary for the definition of ***computer fixed test (CFT) and computer adaptive test (CAT).***

4.1.3. There may be a mechanism in place to review the assessment that are to be potentially administered in a given year, or post-assessment reviews to capitalise on lessons learnt in previous administrations. One of the objectives may be to evaluate the alignment of assessment instruments with the curriculum or target domains. Please specify if this review is regularly conducted by independent qualified experts, is an internal regular process conducted by the ministry or authority in charge of the assessment design, or if those reviews occur on an ad-hoc basis. Please specify if any other mechanism that is practiced in the **yellow cell**, and provide details in the metadata space.

4.2. Characteristics of the items/questions, and background questionnaires

The information requested in this sub-section can be obtained from a sample of latest administered assessment test booklets.

4.2.1. Format of items/questions

Please select the appropriate format:

i. Two options: if there are items or questions to which the answer is to be chosen from a list of two predefined options only.
ii. Three or more options: if there are items or questions to which the answer is to be chosen from a list of three or more predefined options.

iii. Open-ended short constructed response: if there are items or questions where students are granted a score (from a range with a minimum and a maximum) based on a short written answer using their own words.
iv. Essay if students have to elaborate a structured written answer in the form of several paragraphs.
v. Individual project: in technical/vocational assessments the application of knowledge or skills through a specific practical project is generally part of the assessment. Please tick this option if students are to make a presentation to a judging panel or examinations staff about their project, or if they must submit a given output as demonstration that they are capable of applying a certain technique.
vi. Group project: if the project has to be conducted by a group of students.
vii. Other: please specify in the **yellow cell** any other item or question format that is included in the assessment and that is not mentioned among the possible item or question formats listed.

4.2.2. Format of stimulus

The stimulus provides the context to which the student will refer to in order to answer questions or perform certain tasks. Stimuli can have multiple formats. Please select all that apply:

i. Text: if text is used to communicate the context to the student, regardless of the medium used. For example, the text could be provided on paper, or prompted through a screen. Also, text can be presented in various forms including continuous or prose texts, or non-continuous texts such as through lists, tables, schedules, maps etc… Please refer to the glossary for definitions of continuous and non-continuous, mixed and multiple texts.
ii. Audio: if the stimulus is strictly audible. For example, test takers may be asked to listen to a short story and answer some questions, without having the corresponding written text. This content can be played from a recording or delivered in person.
iii. Video: if the stimulus used is a combination of audio and images.
iv. Other. Please describe any other format that is included in the assessment and that is not mentioned among the possible stimuli formats. Provide a short description in the **yellow cell** and any additional details in the metadata space at the end of the section.

 *Please refer to the glossary for the definition of* ***continuous text, non-continuous text, mixed texts and multiple texts.***

4.2.3 - 4.2.4.

Students with special needs: Please refer to the glossary for the definition of autism, hearing impairment, deaf-blindness, deafness, emotional disturbance, physical disability, visual impairment, and speech impairment.

4.2.4 (i) Please tick the type of disability that this assessment accommodates.

4.2.4 (ii) Please describe the types of accommodation(s) that are provided, by type of disability. For example, if the assessment caters to the visually impaired, it may be available in brail, or the stimuli may be presented in a different format – audio as opposed to text.

Please, note that if the option in 4.2.4.i is "No", the appropriate cells in 4.2.4.ii will be automatically turned into code "a" for not applicable.

4.2.5. If test takers with limited proficiency of the language of assessment are accommodated, please describe the type of accommodation(s) in the appropriate cell. For example, test takers may be able to submit their answers to an assessment in another language. This may be common is bilingual settings. Additional time can also be given to them to complete the assessment.

4.2.6. Background questionnaires

In this table you are requested to provide general information regarding the different background questionnaires that are part of the overall assessment. For each questionnaire:

- write the name of the background questionnaire in the first column
- select the group that is responsible for completing the questionnaire. If other, please provide details in the metadata space at the end of the section.
- provide more detail on the official duration for its completion (if you use proxies, please ensure to precede the value by an asterisk)
- elaborate on the intended purpose of this questionnaire. For example the background questionnaire could be completed by parents about their children's engagement in literacy practices at home.

- Provide the nature of data collected. Demographics, household and family structure, education, zone of residence are among the frequent groups of variables used to collect background data.

Example: In PISA 2012, the main assessment test included questionnaires on mathematics, reading, science, problem solving, financial literacy and ICTs. In addition background ***questionnaires*** to be filled in by students, school principals and parents are included which provided the contextual information. This information is generally used for analyses.

 *Please send the latest digital copies of the background questionnaires or test forms for each assessment included in this survey to the UIS. Note that these background questionnaires and test forms will not be made publicly available without prior approval from you or from the organizations in charge of administering the assessment.*

4.2.7. Assessment frameworks

In several national assessments, the data collection tools are elaborated based on specific framework(s) developed by the national assessment team. These frameworks provide the rationale on the structure of the assessment. For example, the assessment frameworks for PISA 2012 are: mathematics framework, reading framework, science framework, problem-solving framework and financial literacy framework.

4.2.8. If available, please provide the list of constructs that are supposed to be measured in this assessment. Generally, they are available in the assessment frameworks.

 Please send the latest digital copies of the *assessment frameworks or lists of constructs covered by the assessment (if they exist) to the UIS. Note that these frameworks will not be made publicly available without prior approval from you or from the organizations in charge of administering the assessment.*

SECTION 5: Coverage and sampling design

5.1. Coverage

This section aims at gathering information about the target population of the assessment, and its geographic coverage.

5.1.1. Geographic coverage: Please select from the pre-defined list the appropriate geographical area covered by this assessment. Note that the assessments targeted by this survey should have national coverage.

5.1.2. Population of interest: Please describe and characterise as much as possible the assessment's target population in order to provide a full definition of the individuals who take this assessment. For example, the target population can be described according to education levels or grades, age, zones of residence (rural, urban), sex, school enrolment status, ownership of schools (public and non-public), special needs status, etc.

5.1.3. Out-of school children

i) Please select **Yes** if out-of school children are included in this assessment, and proceed to 5.1.3 (ii). Otherwise, proceed to item 5.1.4 since the appropriate cells in 5.1.3.ii and 5.1.3.iii will turn automatically into code "a" for not applicable.

ii) Please select **Yes** if participating out-of school children are attending grade equivalent literacy programmes (or second chance programmes) that cover the school curriculum targeted in this assessment (if and when applicable). For example, this can be the case if these children are being sampled from literacy programmes.

iii) Please select **Yes** if passing the assessment allows out-of school children to be re-integrated into the school system at the next appropriate grade.

Note that some public examinations are open to all individuals, whether in or out of school. Passing these assessments may be generally required for entrance into a specific educational programme, regardless of previous school trajectory.

5.1.4. Some groups may be excluded from taking part of the assessment for several reasons that may or may not be directly linked to the objective of the assessment. For example, students with visual impairments may be excluded because the design does not include specific modalities to facilitate their participation. If applicable, please describe what specific groups are officially excluded from this assessment. Other children may be excluded as they reside in areas that hard to access. Please describe the groups excluded from the assessment.

5.1.5. Please select the locations where students are invited to take the assessment, whether in major cities only, local centres across the country (including educational institutions in villages, in small and large cities etc.). For any additional places where students take the test, please specify in the **yellow cell**.

5.2. Participation

When national assessments do not target the whole population of students of a given level, grade or age, a **sample** that is representative of the total target population may be assessed, A representative sample is constructed based on several characteristics or variables such as the number of schools and their distribution across urban/rural areas and major cities, the number of classrooms, the number of candidate students per school/classroom, the age and the sex of candidates. This information is generally available in the technical report of the assessment, and is used later to validate, impute and adjust data before performing analysis.

Please complete both the total target population of students (N) (total number of students in the target population as defined in question 5.1.2) and the total number of test takers (n) per grade covered in this assessment and by sex.

- ✓ Assessments may cover one or more grades. Please select the grade(s) covered (as in 3.2.2). One grade can be selected per row.
- ✓ In the case of a sample-based assessment, total test takers (n) will be the sample size and total (N) will be the number of students identified in the sampling frame of target student population.
- ✓ In the case of a census-based assessment (all students in the target population are assessed), the total (N) will be the number of students enrolled to take the test, while total test takers (n) is the number of students that effectively took the test. In this case, both n and N should be the same, or nearly the same.

In the last column, you are requested to include the age-range of test takers. If possible, differentiate between the target ages, and the actual ages of the test takers. For example you may write "10-15 and 8-9, 16+", and in the metadata space at the end of the section specify that the test targets students aged 10-15 years old, however, since the test is open to all individuals, it also included some aged 8-9 and 16+ years of age.

Please, note that if the option "***Census (all students at the given grade(s) or age level(s)***" was chosen in 3.1.4, then all the cells in sub-sections 5.3 and 5.4 will turn automatically into code "a" for not applicable. In this case, please go directly to the section 6.

5.3. Sampling

5.3.1. Several sampling methods can be used to draw a representative sample of a target population. This however depends on, to name a few, the structure of the sampling frames, the budget, the desired levels of data disaggregation and reporting required for policy making. Particularly in education, samples are drawn at different levels, from administrative areas to students themselves. The most common sampling designs are provided in pre-defined options in question 5.3.1. For the purpose of this survey, please select the sampling design used to draw the administrative units (regions/provinces/states, divisions, sub-divisions, districts, etc.), samples of schools, classrooms and students (if applicable).

5.3.2. Please provide a brief descriptive summary of each sampling frame used to draw the sample(s).

5.3.3. Please provide a brief summary of the design omission if applicable. Design omission is a source of error, referred to as under-coverage. It happens when inaccurate, incomplete or inadequate and out of date sampling frames are used to draw the sample.

Please provide any clarification or additional information that could help to ensure understanding of the sampling methodology in the space provided for metadata.

Please also provide the technical report or any other document that could bring additional information on the sampling methodology.

5.4. **Participation rates and adjustments**
Please refer to the glossary for definitions of the participation rates and adjustment methods before completing this table.

 *Please refer to the glossary for the definition of* ***sample***, ***sampling method***, ***participation rates*** *and* ***weighted participation rates.***

SECTION 6: Data processing

 **6.1. Data processing**

6.1.1. Data editing or data verification: It is the step between data collection and data entry. It is a necessary step that ensures that data is clearly marked on the questionnaire or test booklet before it is captured. Please select the method used to edit data that have been collected, before they are entered. When the paper-pencil version of the assessment is used to collect data, it can happen that **data editing** be performed directly on the paper copies before data are entered in a database. There are also cases where data are entered in a data base, but data editing is not performed through a computer programme, but based on visual control of data. In either of the two cases, please select manual or visual control. But, if a computer programme is used to perform this activity, no matter how the data have been entered in the computer, please select the option **automatic.** If both methods are used in the data editing process, please select both. In the case other data editing methods have been used, please specify the method in the **yellow cell** of the last column.

6.1.2. Data entry/capture:
In general, whether an assessment is computer-based or paper-and-pencil based, data need to be centralised in a database for ease of management (cleaning, querying, analysis, sharing, etc.). Particularly in paper-pencil based test, data available on paper need to be entered carefully in a database, either manually by a trained operator or through a scanning device. In the latter case, scanning transforms paper documents into electronic images or data that can be used in computer-based applications and archives. Recently, scanning is being used more frequently to speed the data processing time of general population and housing censuses, or large scale household surveys, whether the collection is done face-to-face or by mail. Please select whether data capture or entry is manual, scanning or other and specify in the **yellow cell**.

6.1.3. Places for data editing and entry:
Depending on the organisational culture for each assessment, data editing and entry can be performed in local assessment centres available across the country. For example, those can be where students previously went to take the test, in regional centres or at the headquarters of the institution that administers the assessment or has the mandate to perform these activities. Please select the options that apply to this assessment.

 *Please refer to the glossary for the definition of* ***data processing, data editing***, ***data capture*** *and* ***data coding***.

 **6.2. Data appraisal**
Data requested in this sub-section are generally made available to users via technical reports.

6.2.1. & 6.2.2. Please check whether a technical report for this assessment is available. This will facilitate the completion of this section with the appropriate data on the estimation of sampling errors and the actions taken to reduce non-sampling errors. If such a technical report is available online, please provide its URL in the metadata notes at the end of the section.

6.2.3. Candidates are categorised as a non-response if:

i. no data has been collected because they were not reached or they preferred not to take the assessment. Data are totally missing.

ii. the number of test questions/items/tasks completed is not enough (below a minimum to be specified, in percentage of total /questions/items/tasks) to consider their data for further analysis. For example, it may be the case that if respondents attempt 10% of the assessment or less, they are considered as a non-respondent, or as if they did not take the test at all.

If a certain percentage of completion is used as a rule to identify non-respondents, please provide this information in the cell. If another rule is used, please provide it in the **yellow cell** and explain in the metadata.

6.2.4. Item non-response

An item is categorized as non-response (missing data) in several cases. Please select the criteria used in this assessment and specify in the **yellow cell** if another criteria was applied for this assessment but is not listed.

 Include any observations/clarifications that are important to understand and interpret the data in the metadata space at the end of the section.

SECTION 7: Measurements and Results

Please refer to the definitions and examples provided in the instruction manual before completing this section.

7.1. Measurement

7.1.1. Model categories

i) The definitions of CTT, IRT and Rasch models are available in the glossary. Notes on the usages of these models are generally available in the assessment's technical report and may also be introduced in the other reports where students' performances are presented. Please tick all theories of measurement that apply to this assessment, and specify in the **yellow cell** if another type not listed has been used to measure students' performances.

ii) If IRT is chosen above, please specify the model used whether Logistic, Rasch or any other category of models not cited. If the IRT option in 7.1.1.i is "No", the cells in 7.1.1.ii will turn automatically into code "a" for not applicable.

iii) Choose from the list if the results are reported by content domain; as a composite; or by content domain and as a composite at the same time. Reporting by content domain means that results are available per each theoretical knowledge area or subdomain which form together the domain assessed. For example, an assessment in mathematics can include three different subdomains: algebra, geometry and statistics. If results are reported separately for each of those subdomains and there is no global score for mathematics, please choose "reported by content domain". If results are not reported separately for each subdomain, but only as a global score for mathematics, please choose " reported as a composite".

7.1.2. Based on the analytical or technical reports it is possible to identify the grade referencing method used (describing the criteria against which student's performance is compared: either against learning objectives; other students' performance; or his own previous results). Please select all that apply to this assessment and specify in the **yellow cell** any other referencing method not listed but applied in this assessment.

 *Please refer to the glossary for the definition of* ***classical test theory (CTT), item response theory (IRT), Rasch measurement model, norm-referenced method, criterion-referenced method and individual-referenced method.***

7.2. Results

7.2.1. This sub-section should be completed **only if the assessment or public examination described** in section 3.1 **has no streams and/or areas of specialisations**. Otherwise, please skip sub sections 7.2.1 to 7.2.3 and proceed to **7.2.4.**

Please select the metric that applies to the assessment, and provide the minimum score or result that is required for a student to meet the national standard. Please note that when the answer is "No", the corresponding cell for the minimum requirement to meet the national standard turns automatically into code "a" for not applicable.

For example, if the **percentage of correct items** is used, students may be declared to have met the minimum national standard if they answer at least 50% of items correctly, provided that all the items have the same weight. In this case the minimum national standard is 50%.
Generally, when the **average scale score** is used to report on students' performances, the minimum national standard is identified by a raw number. Assuming that the minimum requirement is 500, a student who scores 500 or above is considered to have met the national standard.

Nowadays, usage of **proficiency levels** is becoming more common to reflect and categorise differences in performances of students. Categorising students into groups allows for a target-needs approach. It facilitates putting in place appropriate policies or developing instructional plans in support of groups with specific needs. Both alphabetical letters A, B, C, D, etc. are often used, as well as ranges of numbers. In any case there is an established relation between these two formats. Each letter or range describes the kinds of skills students master or/and lack. If applicable to this assessment, please tick the box for proficiency levels and provide the full explanations or descriptors of each level as reported in the latest version of the official documents of this assessment. Please ensure you provide the descriptions of the proficiency levels in the metadata space at the end of the section.

Note that for international assessments, the metrics and the minimum standard must be those defined by the testing institution, and not those chosen by the participating countries.

7.2.2. In some countries, school-based assessments (tests developed and administered at the school-level) contribute given part or weight of the total final score, particularly for public examinations. If this is the case, please indicate this in the appropriate cell, and provide additional explanation in the metadata. Otherwise (school-based assessments are not collected in this survey), specify 'a' for not applicable.

7.2.3: Percentage of students above the minimum required national standard by gender and by age

i. Some large–scale assessments or public examinations are administered in more than one grade. For this reason, some subjects maybe administered specifically to one grade, or target many grades. There are spaces for four grades. If the assessment covers 4 or more grades, you can insert similar items as in 7.2.3.i-iv (copy rows and insert below 7.2.3.iv). If the assessment or public examination covers only one grade, please choose the option "a" for grade tested in 7.2.3. ii to 7.2.3.iv. In that case, all the cells in these tables will automatically turn into code "a" for not applicable.

ii. First, please write each **mandatory** subject/construct or curriculum area assessed in the appropriate **yellow cells** of the second row. For example it may be that Economics, Mathematics, English, Geography and Accounting are the mandatory subjects tested in a National Public Examination.
If available, write the different constructs or domains that are being measured. While subjects are taught in a specific module and in a standard learning setting, constructs are associated to specific skills or competencies that emerge as result of combining knowledge acquired across the disciplines, personal experience and context. For example, numeracy is not a discipline that is taught per se, but a trait as measured in several assessments. One can also cite the variety of literacies encountered in the 21st century skills literature as examples of constructs.

For example, PISA 2015 assesses four constructs among 15 years old students: *Mathematics literacy, Reading literacy, scientific literacy and Collaborative problem solving skills and competencies.*

For the overall test and per subject/construct tested, please report the percentages of students whose performances are equal or above the minimum required national standard for both male and female, for male only and for female only. In addition, please report the percentages of students whose performances are equal or above the minimum required national standard for each age of the test takers. Please, put the appropriate ages in brackets.

Example: In country X, a large-scale national assessment targets students in grade 6. The corresponding ages of test takers as recorded in the assessment database are 10, 11, 12, 13, 14, 16, 17, 20 and 23 years old.
Officially, only students under the age of 14 are admitted to enrol in grade 6 in this country. But as this is a public examination, even candidates not currently enroled in schools can take the test, so long as they are qualified and have enroled. For this specific case, it makes sense to report data separately for 10, 11, 12, 13 and 14 years, and for 15 and above. In case you use such a range of ages, please provide explanations in the metadata that will inform the user of these data.

In each table, you must also indicate the language in which each of the above listed **mandatory** subjects or constructs are tested, the official duration of the test for each **mandatory** subject/construct, and the weight in percentage each domain covered carries to the overall score of the subject/construct. For example, consider the example Mathematics includes 4 domains: Algebra, Geometry, Probability and Statistics. Algebra accounts for 25% of the total grade, Geometry 30%, Probability 15% Statistics 30%.

When appropriate, please also provide the weight of each subject/construct tested as part of the assessment or the examination, as the percentage of the total average score performance of the test taker. This is usually the case in public examinations. For example in Mauritania, the CEPAS/CEF is a public examination organised each year by the Ministry of National Education. It is administered on voluntary basis to all students at the end of primary education (ISCED 1), or grade 6. It is comprised of seven mandatory subjects listed below, and the language of the test is either Arabic or French. The titles of each subject, the time allocated in minutes, the language of testing, and the weights as percentage of total average score to this examination for each mandatory subject are provided below:

Subject	Time allocation (mn)	Language of test	Weight (% average total score)
Arabic	90	Arabic	25
Mathematics	90	Arabic	25
French	60	French	15
History and geography	45	Arabic	10
Natural sciences	45	French	10
Religion	45	Arabic	10
Civics	30	Arabic	5

The results of students assessments can also be reported only by subject or domain tested, without having the aggregate score. That is the case in some public examinations where students should take a list of mandatory subjects, as well as optional subjects to get a certification. In that case, please only report the weight per domain covered as percentage of total score in the subject/construct, if available, while the last row should be filled with the code “a” for not applicable.

7.2.4: This sub-section should be completed only if the assessment described in section 3.1 has streams and/or areas of specialisation (to be filled in the cells).

i. First, please write each stream in the appropriate cells.

ii. Then, for each stream, choose the appropriate metrics used, and provide the minimum required to meeting the national standard, the weight of school-based assessments, and the list of areas of specialisation available within the stream.

7.2.5: The streams provided above appear automatically in the first row of this table.
Please complete this table only if the streams have no areas of specialisation.

For each stream (but not subjects tested in the stream), please provide stream aggregated results for both male and female, male only, female only and by ages.

In the last row of this table, please write the list of mandatory subjects tested per stream.

7.2.6: **Uses of the results**
The most common uses of assessments are listed in table 7.2.6. Please select the intensity (only one option per row) in which the assessment **has been used** and include any additional information on the use of assessment results in the column for comments or in the metadata section. Please illustrate each relevant use with an example, whenever possible. This is very important as a section in the country-reports has been dedicated to the impact of the assessment on policy-making. This is different from purpose: an assessment’s purpose is the end(s) for which it has been designed, and this tends to be explicitly stated by the proponents of the assessment early on in the process; an assessment’s uses may go beyond its original purpose, as other actors have access to the data from the assessment and utilize them for their own goals.

In the metadata space, please provide any additional information that will help complete and understand the data requested in this section of the template.

SECTION 8: Data dissemination, reporting and accessibility

8.1. Reporting level

To make data relevant for policy making, monitoring and evaluation, analysis needs to be made at different administrative unit levels, and different population groups. For example, gender sensitive analysis is required to monitor the gender gap between male and female students. Please tick all the levels at which the results of the assessment are reported, and provide additional information in the metadata at the end of the section.

8.2. Data dissemination

Many platforms can be used to inform the public and the education community about the results and data availability of an assessment. Please tick all the platforms used to disseminate data for the target assessment.

8.3. Data accessibility

Micro data is not easily accessible, particularly for student assessments. This sub-section seeks to collect all practical information that would need to be known by researchers or by anyone interested using the micro-data and related documentations for further research. This information is not always documented. Please rely on officials to collect such information or to validate the initial information that may be available to make sure that it is still applicable, and complete the appropriate cells of the template.

8.4. Challenges

The administration and management of large-scale assessments are always challenging, particularly in developing countries. These challenges are generally reported for future considerations. Please provide a short summary of these challenges in the appropriate space regarding the following phases of the students' assessment process: planning, test design, data collection, data processing, data analysis, educational policy reform, data dissemination and other.

*Please refer to the glossary for the definition of **data dissemination.***

5. GLOSSARY AND DEFINITIONS

Assessment of learning outcomes refers to a measure of individuals' achievement of learning objectives. They can focus on specific curriculum areas, use a variety of assessment methods (written, oral and practical tests/examinations, projects and portfolios), and be administered during or at the end of an educational programme.

Assessment types:

National assessment: For the purpose of the UIS catalogue survey, it refers to an assessment of students learning outcomes which aims to describe the achievement of students at a particular age or grade level and provide feedback on a limited number of outcome measures that are considered important at the level of the education system by policy makers, politicians, and the broader educational community. It is generally administered to a sample of students and also collects background information (from students, teachers and parents) that are important to link analysis to policy questions at the national, subnational and local levels.

International assessment is an assessment of learning outcomes that provides similar information as the national assessment, but for more than one national education system. It is generally not sensitive to individual systems since its main goal is the comparability of the results among the participating countries.

Public examinations: For the purposes of the UIS catalogue, it designates exit or end-point standardized exams that are generally set by a central federal/state examining board of a given country in order to promote, select or provide a certification to **all** candidates who qualify or are supposed to have formally or informally learned and covered the curriculum of a formal education programme, as part of their requirements for graduation. The public examination is generally administered each year, to everybody who registers and regardless of age and occupation (in some countries). As opposed to national assessments, students background data is rarely collected along with the public examinations.

Candidate non-response is a status of respondents that occurs when selected participants have not completed the questionnaire or when the completion rate of the questionnaire is not enough to further consider their data for analytical purposes.

Census: An official survey involving the whole population within a defined system. For example, a school census involves all the schools within the education system. For the purpose of UIS catalogue, the target population is either students of a given age or grade level, or enroled candidates who fulfill conditions to pass a given public examination, or to participate to an assessment.

Competence: The ability to mobilize and use internal resources such as knowledge, skills and attitude, as well as external resources such as databases, colleagues, peers, libraries, instruments etc., in order to solve a specific problem efficiently in real life situations.

Computer adaptive testing (CAT) is a method of administering tests that adapts the choice of items to respondent's levels of attainment, as determined by previous responses. A CAT model is very different to a computer fixed test, as it selects items based on prior responses for each respondent, and can record the way each respondent searches for, stores and retrieves information.

Computerized fixed test (CFT) is a digital version of a paper-pencil test, which aims to facilitate data processing and reduce the likelihood of human error when compiling a database.

Cut-score: refers to a particular point on the score scale of a test, such that scores at, below and above it are interpreted differently.

Data capture: refers to the process that consists of entering data in a database system, particularly when the data collection was not computer-assisted.

Data coding refers to the process of assigning numerical values to responses that are originally in a given format such as numerical, text, audio or video. The main objective is to facilitate the automatic treatment of the data for analytical purposes.

Data dissemination refers to the release of data to users through various media (new media and traditional media) such as internet or online media, press conference or release, article in print newspaper, television or radio interview, etc...

Data editing refers to checks which aim to identify inconsistent, invalid and missing data records in order to correct or replace them through a follow-up with the respondent, a manual review of the data or an imputation.

Data processing is a series of manual, automatic or electronic operations such as validation, sorting, summarization, aggregation of data. These operations are usually followed with data retrieval, transformation, classification, data analysis and reporting.

Disabilities: an umbrella term, covering impairments, activity limitations, and participation restrictions. Impairment is a problem in body function or structure; an activity limitation is a difficulty encountered by an individual in executing a task or action; while a participation restriction is a problem experienced by an individual in involvement in life situations. It is a complex phenomenon, reflecting the interaction between features of a person's body and features of the society in which they live. Overcoming the difficulties faced by people with disabilities requires interventions to remove environmental and social barriers (World Health Organization[1]).

> **Autism**: means a developmental disability significantly affecting verbal and nonverbal communication and social interaction, generally evident before age three that adversely affects a child's educational performance. Other characteristics associated with autism are engaging in repetitive activities and stereotyped movements, resistance to environmental change or change in daily routines, and unusual responses to sensory experiences. The term autism does not apply if the child's educational performance is adversely affected primarily because the child has an emotional disturbance. A child who shows the characteristics of autism after age 3 could be diagnosed as having autism if the criteria above are satisfied (US Congress, IDEA 1997).
>
> **Deaf-blindness**: means concomitant [simultaneous] hearing and visual impairments, the combination of which causes such severe communication and other developmental and educational needs that they cannot be accommodated in special education programs solely for children with deafness or children with blindness (US Congress, IDEA 1997).
>
> **Emotional disturbance**: means a condition exhibiting one or more of the following characteristics over a long period of time and to a marked degree that adversely affects a child's educational performance: An inability to learn that cannot be explained by intellectual, sensory, or health f actors- An inability to build or maintain satisfactory interpersonal relationships with peers and teachers- Inappropriate types of behavior or feelings under normal circumstances- A general pervasive mood of unhappiness or depression- A tendency to develop physical symptoms or fears associated with personal or school problems- The term includes schizophrenia. The term does not apply to children who are socially maladjusted, unless it is determined that they have an emotional disturbance (US Congress, IDEA[2] 1997).
>
> **Environmental effect**: Where students who are unable to cope with the assessment environment are accommodated. For example, they may be given an alternative test, or the chance to write it at a different time.
>
> **Hearing impairment (**including **deafness):** impairment in hearing, whether permanent or fluctuating, that adversely affects a child's educational performance. Deafness refers to a hearing impairment so severe that a child is impaired in processing linguistic information through hearing, with or without amplification that adversely affects a child's educational performance (US Congress, IDEA 1997).

[1] http://www.who.int/topics/disabilities/en/

[2] **IDEA:** Individuals with Disabilities Education Act (US Congress, 1997)

Speech or language impairment: means a communication disorder such as stuttering, impaired articulation, language impairment, or a voice impairment that adversely affects a child's educational performance (US Congress, IDEA 1997).

Visual impairment (including **blindness)**: means impairment in vision that, even with correction, adversely affects a child's educational performance. The term includes both partial sight and blindness (US Congress, IDEA 1997).

Format of the stimulus[3]:

Continuous text Continuous texts are formed by sentences organised into paragraphs. Examples of continuous texts include newspaper reports, essays, novels, short stories, reviews and letters.

Non-continuous text: Non-continuous texts, also known as documents, are organised differently than continuous texts, and therefore require a different kind of reading approach. Examples of non-continuous text objects are lists, tables, graphs, diagrams, advertisements, schedules, catalogues, indexes and forms. These text objects occur in both fixed and dynamic texts.

Mixed texts: Mixed texts are ccoherent objects consisting of a set of elements in both a continuous and non-continuous format. In well-constructed mixed texts the components (for example, a prose explanation including a graph or table) are mutually supportive through coherence and cohesion links at the local and global level. This is a common format in magazines, reference books and reports, where authors employ a variety of presentations to communicate information.

Multiple texts: Multiple texts are texts that have been generated independently, and make sense independently; they are juxtaposed for a particular occasion or may be loosely linked together for the purposes of the assessment. The relationship between the texts may not be obvious; they may be complementary or may contradict one another. For example, a set of websites from different companies providing travel advice may or may not provide similar directions to tourists. Multiple texts may have a single "pure" format (for example, continuous), or may include both continuous and non-continuous texts.

Grade: a specific stage of instruction in initial education usually covered during an academic year. Students in the same grade are usually of similar age. This is also referred to as a 'class', 'cohort' or 'year'(UIS glossary).

Grade referencing method

Criterion-referenced assessment (CRA): When students are tested against a pre-defined standard of performance, or target, or desirable performance, benchmark, or criterion, the assessment is criterion-based. In CRAs, all students are assessed against the criterion, or reference, which could be a specific body of knowledge and skills. In education, CRAs usually are made to determine whether a student has mastered the material taught in a specific grade or course.

Individual/Ipsative assessment: consists in assessing present performance of a person against his prior performance, with a view to determine if any improvement has been made. This practice requires longitudinal data.

Norm-referenced assessment (NRA): refers to the process of comparing one test-taker to his or her peers, i.e. when students' scores are simply ranked from low to high, and each student's ranking is compared to the rankings of others. There is no attempt to interpret the scores in terms of what students know and can do, except in the limited sense that a student's performance is typical of other low, middle, or high performing students in the group.

Typical individual results statement: *The student obtained a scale score of 450 (on a scale with mean 500, standard deviation 100).*
Typical group results statement: *The average scale score for grade X in assessment Y is S and the standard deviation is D.*

[3] OECD 2015 PISA reading framework, pp 17-18.

Notes: 1-In a norm-referenced test, the scores are often recast as percentiles: a student with score X is given the percentile P where P% of the students have scores of X or less. In this way, it is seen that NRTs are designed to sort and rank students "on the curve," not to see if they meet a standard or criterion. Therefore, NRAs should not be used to assess whether students have met desirable (or a minimum) standards. 2- A test can in theory be both criterion- and norm-referenced. A CRA may be used to measure student learning in relation to standards, with specific cut-off scores on the scale chosen to separate levels of achievement on the standards. And then the scores can also be ranked. But it might not have sufficient accuracy for NRA purposes. A NRA may be evaluated for use as a CRA, but it might not have sufficient content and accuracy for determining levels of knowledge and skills. A test should only be used for both objectives when these dual objectives are clearly defined during test development.

Percentage of correct items: The number of test items that a student answers correctly, divided by the total number of test items, times 100 is that student´s percentage score on that test. That is, the score is a percentage rather than the number of correct responses. If a test has items with several score points, then the student's percentage score is the total score points earned over all items, divided by the total possible number of score points to be earned, times 100.

Typical individual result statements: The student answered X% of the test items correctly. The student earned X% of the possible points on the test.
Typical group results statements: The average percentage correct in the group was X% with a standard deviation of Y%. The percentage of points earned ranged from X% to Y% with average of Z%.

ISCED: A classification system that provides a framework for the comprehensive statistical description of national educational systems and a methodology that translates national educational programmes into internationally comparable levels of education. The basic unit of classification in ISCED is the educational programme. ISCED also classifies programmes by field of study, programme orientation and destination. Please click the link below for more information on the ISCED classification.

http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

Item: refers to a single question or task of an assessment test.

Note: An item has three parts: the stimulus that provides the context of the task in audio, video or text format, a clear description of the task to be performed based on the stimuli, and the potential response options in a given format (dichotomous, multiple choices, partial credit, open-ended, essay, etc.)

Mandatory assessment refers to a ***high-stake assessment*** that is administered to all students of a given grade. For the purpose of this survey, the General Certificate of Education Ordinary level (GCE-OL) or Advanced level (GCE-AL) in the British system, the Baccalaureate (Bacc) in the French system, the Diplôme d'étude collégiale (DEC) in Quebec-Canada are examples of mandatory assessments. Only students having the GCE-AL, the Bacc or the DEC or equivalences are allowed to enroll at higher education or to take some test for selection to upper secondary schooling.

Matrix sampling design is the process by which different subsets of test items are administered to different subsets of respondents, in a way that each item is administered to at least one subset of the sample of respondents. Generally, there are important items called 'core' that are administered to all respondents, as opposed to 'split' items that are only administered to a subset of respondents. In general, core items are predictive of many of the split items of the test.

Measurement:

Classical Test Theory (CTT): A measurement theory which consists of a set of assumptions about the relationships between actual or observed test scores and the factors that affect the scores. It is used for measuring and managing test and item performance data. In contrast to item response theory, it comprises a set of more traditional psychometric

methods (DFID[4]). In CTT, the global level ability of a test taker depends on the overall test difficulty and vice-versa. There is no way to obtain parameters of each test item or the level ability of the test taker in each item on a scale performance continuum as in IRT and Rasch models.

Rash measurement model: A group of mathematical models for relating and predicting the individual's score on a test item to his level of performance on a scale of the ability or trait being measured, and the item's difficulty parameter. The Rasch model considers that the probability of an individual to provide the right answer or to perform efficiently a given task only depends on the difference between his level of ability and the level of item difficulty. This probability increases when the ability is higher than the item difficulty, but is 50% if these two parameters are equal.

Item Response Theory (IRT): A group of mathematical models for relating and predicting the individual's performance on a test item to his level of performance on a scale of the ability or trait being measured, and the item's characteristic parameters: guessing, discrimination and difficulty parameters.

Note: *Compared to the classical test theory (CTT), both IRT and Rasch modelling have the potential to provide estimate of the level ability of test taker that are independent of the test items, and the estimates of the item characteristics that are also independent of the actual group of test. This is called the invariance property.*

Non-response adjustment method is a statistical method used to replace missing or invalid data. Weighting and imputation are examples of non-response adjustment methods generally used to account for non-responses before analyses are performed.

Imputation is a statistical procedure used to determine replacement values to missing, invalid or inconsistent data. Examples of frequently used imputation methods are: hot deck, cold deck, listwise and pairwise deletion, mean imputation, regression imputation, last observation carried forward, stochastic imputation and multiple imputations.

Participation rate: Is the total number of individuals in the selected sample who have participated to the survey (completed the survey questionnaire) expressed as the percentage of the initial number of selected individuals (sample size) or total target population.

Pilot survey: a preliminary survey that is conducted with few individuals of the target population or the sample of a survey, in order to test and refine the survey instruments (questionnaire and instruction manual, data processing manual and programmes) before the main data collection across the target population or the full sample.

Plausible values are ability estimates generated through a combination of item response modeling techniques and multiple imputations from the latent regression models. Even though plausible values are not real observed test score, they have the property of a test score that allow secondary analysts to use standard software and techniques to analyze data that have been collected using complex matrix sampling designs.

Proficiency levels: It refers to the classification of students into bands of performance that are identified by series of cut-off scores on the performance scale. Proficiency levels are commonly used in criterion-referenced tests. For each of these levels, there should be descriptions of what it means to be in the level, in terms of knowledge, skills, attitudes, etc. Each level (or band) represents a degree of mastery of what the test purports to measure. The same levels of proficiency can be expressed with words or letters. For example, "below basic", "basic", "above basic"; or high, middle, low; or A, B, C.

Typical individual results statement: The student is classified into level 2 or Basic. The student is classified into level 4 or Advanced.

[4] http://www.heart-resources.org/wp-content/uploads/2012/04/04-Learning-Outcomes-How-To-Note.pdf

Typical group results statement: In schools of this type, X% fall in level 1 (Below Basic), Y% fall in level 2 (Basic), Z% fall in level 3 (Proficient), and W% fall in level 4 (Advanced). In this school, X% of the students are found in level 2 (Proficient) or above.

Note: Proficiency levels are usually found in conjunction with proficiency scale scores. Determining cut-off scores for a particular test scale starts with definitions and descriptions of what is meant by the proficiency levels in terms of knowledge and skills, evaluates the items along the scale for content and complexity, and then proceeds with systematically determined judgments about where to put the cut-scores.

Reporting metrics refers to the different forms of reporting the results of assessment of student learning outcomes, or student achievement, for communication purposes with the education stakeholders including the students. The results can be reported for individuals or aggregated for specific groups. The possible forms (metrics) of reporting results include percentage pass or fail, scale and proficiency levels.

Sampling design

Sample is a subset of individuals from a specific population, formed according to a selection process (with or without replacement and at random or not). A sample that is formed by random selection of individuals based on known probabilities is called random or probability sample. A non-random sample is formed on the basis of subjective method of selection.

Sampling is a method of designing a sample based on a (sampling) frame that contains all individuals of the target population. Methods of sampling a population depends on many parameters such as the sample selection process (with or without replacement and at random or not), the structure of sampling frame, the level of data disaggregation and analysis needed, and the budget available for the study. Below are common probability and non-probability sampling methods (Statistics Canada)[5]:

Probability sampling:

Simple random sampling: each member of a population has an equal chance of being included in the sample. For that to happen, all of the units in the survey population should be listed.

Systematic or interval sampling, means that there is a gap, or interval, between each selected unit in the sample. In order to select a systematic sample, you need to follow these steps:

i. Number the units on your frame from 1 to N (where N is the total population size).
ii. Determine the sampling interval (K) by dividing the number of units in the population by the desired sample size.
iii. Select a number between one and K at random. This number is called the random start and would be the first number included in your sample.
iv. Select every Kth (in this case, every fourth) unit after that first number.

Sampling with probability proportional to size: Probability sampling requires that each member of the survey population have a chance of being included in the sample, but it does not require that this chance be the same for everyone. If information is available about the size of each unit (e.g., number of students for each school or classroom) and if those units vary in size, this information can be used in the sampling selection in order to increase the efficiency. This is known as sampling with probability proportional to size (PPS). With this method, the bigger the size of the unit, the higher the chance it has of being included in the sample. For this method to bring increased efficiency, the measure of size needs to be accurate.

[5] http://www.statcan.gc.ca/edu/power-pouvoir/ch13/prob/5214899-eng.htm#a1
http://www.statcan.gc.ca/edu/power-pouvoir/ch13/nonprob/5214898-eng.htm

Stratified sampling: the population is divided into homogeneous, mutually exclusive groups called strata, and then independent samples are selected from each stratum. Any of the sampling methods can be used to sample within each stratum, and the sampling method can vary from one stratum to another. When simple random sampling is used to select the sample within each stratum, the sample design is called stratified simple random sampling. A population can be stratified by any variable that is available for all units on the sampling frame prior to sampling (e.g., student grades, sex, age, province/region/zone of residence, school ownership).

Cluster sampling: The population is first divided into groups or clusters. A number of clusters are selected randomly to represent the total population, and then all units within selected clusters are included in the sample. No units from non-selected clusters are included in the sample—they are represented by those from selected clusters. This differs from stratified sampling, where some units are selected from each group.

Multi-stage sampling is like the cluster method, except that it involves picking a sample from within each chosen cluster, rather than including all units in the cluster. This type of sampling requires at least two stages. In the first stage, large groups or clusters are identified and selected. These clusters contain more population units than are needed for the final sample. In the second stage, population units are picked from within the selected clusters (using any of the possible probability sampling methods) for a final sample. If more than two stages are used, the process of choosing population units within clusters continues until there is a final sample.

Multi-phase sampling:

A multi-phase sample collects basic information from a large sample of units and then, for a subsample of these units, collects more detailed information. The most common form of multi-phase sampling is two-phase sampling (or double sampling), but three or more phases are also possible.

Multi-phase sampling is quite different from multi-stage sampling, despite the similarities in name. Although multi-phase sampling also involves taking two or more samples, all samples are drawn from the same frame and at each phase the units are structurally the same. However, as with multi-stage sampling, the more phases used, the more complex the sample design and estimation will become.

Non-probability sampling[6]:

Convenience/ haphazard or accidental sampling: sample units are only selected if they can be accessed easily and conveniently.

Volunteer sampling: This type of sampling occurs when people volunteer their services for the study.

Judgment sampling: A sample is taken based on certain judgments about the overall population. The underlying assumption is that the investigator will select units that are characteristic of the population.

Quota sampling: Sampling is done until a specific number of units (quotas) for various sub-populations have been selected. Since there are no rules as to how these quotas are to be filled, quota sampling is really a means for satisfying sample size objectives for certain sub-populations.

Out-of-school children: any children of primary or lower secondary school age who are not enrolled in primary or secondary education are considered to be out of school. This includes a small number of children in pre-primary education and in non-formal education. Children of primary school age who are enrolled in pre-primary education are counted as out of school, because the educational content of pre-primary education and the pedagogical qualifications of its teaching staff are not equivalent to the standards required for primary education. Children in non-formal education programmes are also considered

to be out of school, because the nature of these programmes is not, in general, equivalent to that of formal primary and lower secondary education (UIS and UNICEF[7]).

Scale scores: Scale scores are ability estimates from the Item Response Theory (IRT) models, based on students' response vectors. Scale scores are designed to provide a metric that is consistent for different versions of a test and consistent across time.

School-based assessment refers to student assessment regularly organized and administered by each educational institution established in a country. Assessment tools are generally designed by the teachers' staff of the institution. The results are used to provide direct feedback to students and parents, to regulate classroom and improve the teaching-learning process. In some countries, scores to these assessments count (weight on the final total score) for the graduation or selection of students.

Stake of the assessment

> **High stake assessment**: A high-stake assessment is an assessment with important consequences for the **test taker**, on the basis of his performance. Passing has important benefits, such as progressing to a higher grade, a high school diploma, a scholarship, entrance into the labor market or getting a license to practice a profession. Failing also has consequences, such as being forced to take remedial classes or not being able to practice a profession. Examples of high-stakes tests include college entrance examinations, high/secondary school exit examinations, and professional licensing examinations
>
> **Low stake assessment** is meant to verify how well the education system (or school system) is doing.
> Low-stakes assessments have no direct consequences on test takers. For example, regardless of their scores, students will not be held in a grade, or be denied access to higher levels of schooling, or be streamed into more academic and less academic programs.
> However this neutrality may not apply to all involved parties. There may be high stakes for teachers and schools if those with poorly performing students are singled out by the proper authorities for various actions (firing, retraining, etc.) Government administrations may find themselves under political fire by opposition parties when system results are low or decline. Hence, while for the test takers, the students, the assessment may be low-stakes, it may become high-stake to others.

Standardised test: A test in which items/tasks or questions, administration conditions, editing, scoring and interpretation of results are applied in a consistent and pre-determined manner for all test-takers.

Trait/Construct is a hypothesised unobservable or mental trait that is used to explain individuals' performance on an assessment. It is only measured trough observations or tasks performances from which the level of the test taker is inferred. Constructs cannot necessarily be directly observed or measured.

Weighted participation rate is a sample specific participation rate that accounts for the weight of each individual who participated to the survey.

[7] http://www.uis.unesco.org/Education/Documents/oosci-global-report-en.pdf