The Data Quality Assurance Framework

A framework to support the monitoring of learning outcomes for Sustainable Development Goal 4

Concept Note

UNESCO
United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
for
STATISTICS

ACER

Table of contents

The Data Quality Assurance Framework 2
Concept Note 2
About this document 1
Section 1: Overview of the proposed UIS approach to monitoring learning for SDG 4 2
Section 2: The International Code of Practice in Learning Assessment 6
  Why develop an ICP-LA? 6
  Situating the ICP-LA in the broader SDG framework for data quality 7
  Structure and content of the ICP-LA 9
  Style of the ICP-LA 14
  Supporting the interpretation and application of the ICP-LA 14
Section 3: The Evaluation of Alignment Process 16
  Alignment within the Evaluation of Alignment Process 16
  Proposed Evaluation of Alignment Process 16
    Domain alignment 17
    Strand alignment 18
    Level alignment 20
    Method 21
Section 4: The Assessment of Data Process 27
  Introduction to the ADP 27
  ADP tools and methods 28
  Auditing, adjudicating and reporting ADP decisions 30
  The relationship between the ADP and other initiatives that explore the quality of learning assessment systems and products 31
Section 5: Next steps 32
Bibliography 35

About this document

This document presents the Data Quality Assurance Framework (DQAF), which is part of the proposed approach by UNESCO's Institute for Statistics (UIS) to monitoring learning outcomes for the UN's Sustainable Development Goal 4 (SDG 4). It has been prepared for the second meeting of the Global Alliance to Monitor Learning (GAML), in Washington DC on 17–18 October 2016.

This document lays out the work to date of UIS and its technical partner ACER-GEM, and is intended to guide further work on the DQAF that is undertaken within the GAML network or by external associates of GAML. The content of each section of this document is summarised below.

- *Section 1* gives an overview of the approach UIS is advocating for monitoring learning outcomes for SDG 4.
- *Section 2*, discusses the International Code of Practice in Learning Assessment (ICP-LA), the first of the three elements of the DQAF.
- *Section 3* discusses the Evaluation of Alignment Process (EAP), the second of the three elements of the DQAF.
- *Section* 4 discusses the Assessment of Data Process (ADP), the third of the three elements of the DQAF.

  Throughout *Section 2 – Section 4*, some issues for further discussion are raised. They are indicated as shown below.

  > **Issue for further discussion**
  >
  > Issues about the features and characteristics of the DQAF that require further discussion within the GAML network are in shaded boxes like this one.

- *Section 5* discusses how the issues raised in Section 2 – Section 4, and some more general questions, might be addressed by the GAML network so that the development of the DQAF can progress.

Section 1: Overview of the proposed UIS approach to monitoring learning for SDG 4

Sustainable Development Goal 4 (SDG 4) is to 'ensure inclusive and equitable quality education and promote lifelong learning opportunities for all'. Tracking progress towards some of the targets associated with this goal will require the measurement of learning outcomes at several different stages of life.

The education sector faces challenges in measuring learning outcomes for this goal because it is not possible to establish narrow and unambiguous definitions of key indicators, nor is it possible to implement a single measurement process with broad cross-national applicability. This is because concepts of learning and how it should be measured are dependent on local context, and countries make independent interpretations and decisions about these matters. This does not mean that international comparability of learning outcomes is out of reach, nor does it mean that there is no place for statements about how learning should be measured and reported. Rather, it means that fit-for-purpose international comparability must be achieved without the imposition of universal measurement processes. It also means that, if they are to be as useful as possible, statements about how learning should be measured and reported – that is, statements about best practice in the field – must cover a variety of approaches.

UNESCO's Institute for Statistics (UIS), through its Global Alliance to Monitor Learning (GAML), is working on an approach to monitor learning outcomes for SDG 4. The UIS reporting scales are a central part of this work. The UIS reporting scales are numerical scales and associated substantive descriptions that explain developing proficiency in the learning domains that feature in the SDG 4 targets. The drafting of the scales has been informed by items and data from a number of different learning assessments.[1]

Particular locations on the UIS reporting scales will be established as benchmarks. Through the benchmarks and the substantive descriptions of developing proficiency, the UIS reporting scales will provide a backbone for shared interpretations of the SDG 4 indicators that refer to learning outcomes. For example, global indicator 4.1.1 is:

> Percentage of children/young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics

(Inter-Agency and Expert Group on Sustainable Development Goal Indicators, 2016)

[1] Scales have initially been developed for literacy/reading and numeracy/mathematics. The possibility of developing scales in other learning domains (e.g. global citizenship and digital literacy) will be explored at a later date.

The substantive descriptions on the UIS reporting scales will provide a backbone for interpreting the words ‘reading’ and ‘mathematics’, and the benchmarks a backbone for interpreting the expression ‘at least minimum proficiency’ for each of grades 2/3, end of primary and end of lower secondary.[2]

Tools and methods will also be developed to link results from different assessments so they can be reported against the UIS reporting scales. A Data Quality Assurance Framework (DQAF) will also be prepared. This framework will enable the quality and appropriateness of learning assessment data submitted for reporting against the UIS reporting scales to be judged and assured. The DQAF will be informed by an International Code of Practice in Learning Assessment (ICP-LA).

Together, the interpretive backbone of the scales, the linking tools and methods, and the DQAF will enable countries to measure and report on learning in ways that are appropriate for their different contexts, yet still facilitate fit-for-purpose international comparisons, and encourage and promote best practice in learning assessment.

The entire process for reporting against the UIS reporting scales is described below, and illustrated in Figure 1.

1. A country approaches UIS and indicates that it wishes to submit results from an assessment for reporting against the UIS reporting scales.
2. The Evaluation of Alignment Process (EAP) is undertaken by the responsible organisation in conjunction with assessment experts from the country. In this process, available information from the country’s assessment (eg assessment frameworks or test blueprints, sample items, described proficiency scales) is analysed to determine the extent to which there is overlap between the knowledge, skills and understandings that are tested in the assessment and the knowledge, skills and understandings that feature in the substantive descriptions on the UIS reporting scales. Note it is not expected that any one assessment will cover the entire span of the UIS reporting scales, but rather that assessments targeted to different age/grade levels will cover different ranges of the UIS reporting scales.
    - ➢ If the EAP determines that there is adequate overlap between the knowledge, skills and understandings tested by the assessment and the knowledge, skills and understandings that feature in the substantive descriptions in the UIS reporting scales, then the country’s assessment will proceed to the next step, the Assessment of Data Process (ADP) – *go to step 3 below*.

[2] Note that the substantive descriptions on the UIS reporting scales provide interpretive backbones for the learning domains, not rigid and narrow definitions of the learning domains. In other words, the substantive descriptions are sufficiently broad and detailed that they will be able to accommodate the variations in interpretation of the learning domains that exist across countries.

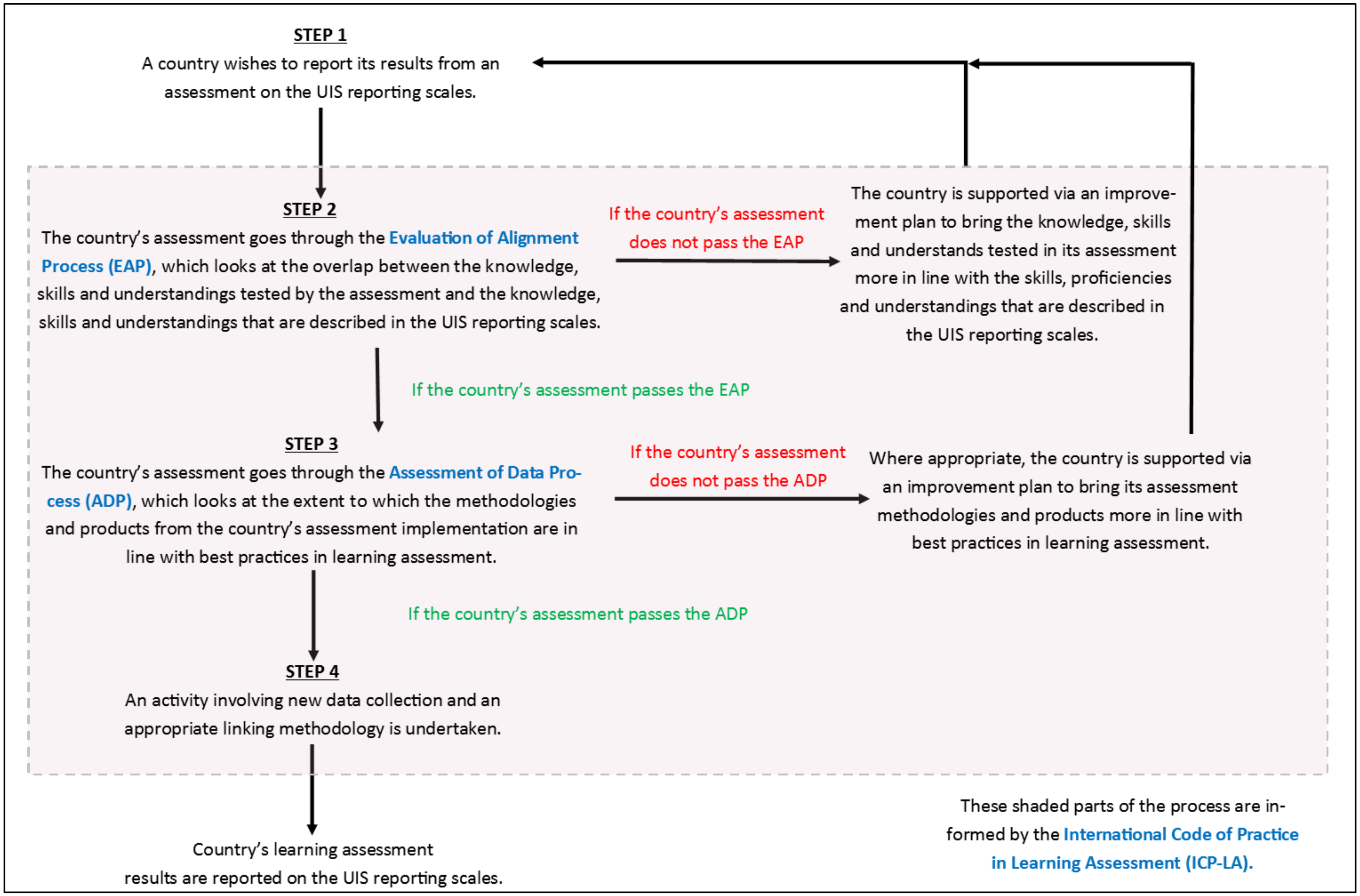
   - If the EAP determines that there is not adequate overlap between the knowledge, skills and understandings tested by the country's assessment and the knowledge, skills and understandings that feature in the substantive descriptions in the UIS reporting scales, then the country will be supported via an agreed improvement plan to adjust its assessment so there is improved overlap between the knowledge, skills and understandings it tests that the knowledge skills and understandings that feature in substantive descriptions in the UIS reporting scales. In subsequent assessment cycles, after the improvement plan has been enacted, the country may wish to again submit results from its assessment for reporting against the UIS reporting scales – *return to step 1 above*.

3. The Assessment of Data Process (ADP) is undertaken by the responsible organisation in conjunction with assessment contacts in the country. In this process, the methods and products from the country's implementation of the assessment are examined to determine the extent to which they are in line with best practices in learning assessment, as articulated in the ICP-LA. This examination will involve, for example, document review, interviews, observations and inspection of databases.

   - If the ADP determines that the methods and products from the country's implementation of the assessment are in line with best practice in learning assessment as articulated in the ICP-LA, then the country's assessment will proceed to the linking step – *see step 4 below*.

   - If the ADP determines that the methods and products from the country's implementation of the assessment are not sufficiently in line with best practice in learning assessment as articulated in the ICP-LA, then the country will be supported via an agreed improvement plan to modify its methods and products to bring them more in line with the ICP-LA. In subsequent assessment cycles, after the improvement plan has been enacted, the country may wish to again submit results from its assessment for reporting against the UIS reporting scales – *return to step 1 above*.

4. An appropriate linking activity involving new data collection is undertaken. In this activity, items from the country's assessment and a selection of items from assessments that informed the development of the UIS reporting scales will be administered to children in order to obtain a fit-for-purpose mapping of the country's assessment results onto the UIS reporting scales.



STEP 1

A country wishes to report its results from an assessment on the UIS reporting scales.

STEP 2

The country's assessment goes through the Evaluation of Alignment Process (EAP), which looks at the overlap between the knowledge, skills and understandings tested by the assessment and the knowledge, skills and understandings that are described in the UIS reporting scales.

If the country's assessment does not pass the EAP

The country is supported via an improvement plan to bring the knowledge, skills and understands tested in its assessment more in line with the skills, proficiencies and understandings that are described in the UIS reporting scales.

If the country's assessment passes the EAP

STEP 3

The country's assessment goes through the Assessment of Data Process (ADP), which looks at the extent to which the methodologies and products from the country's assessment implementation are in line with best practices in learning assessment.

If the country's assessment does not pass the ADP

Where appropriate, the country is supported via an improvement plan to bring its assessment methodologies and products more in line with best practices in learning assessment.

If the country's assessment passes the ADP

STEP 4

An activity involving new data collection and an appropriate linking methodology is undertaken.

Country's learning assessment results are reported on the UIS reporting scales.

These shaded parts of the process are informed by the International Code of Practice in Learning Assessment (ICP-LA).

Figure 1: The process for reporting assessment results against the UIS reporting scales

Section 2: The International Code of Practice in Learning Assessment

This section first argues for the need of an ICP-LA, then situates the ICP-LA in the broader context of UN principles about the quality of statistics, then presents some general notes about the envisaged content, structure and style of the ICP-LA, and finally discusses some supplementary information to support the interpretation and application of the ICP-LA.

Why develop an ICP-LA?

There are already many different statements of principles and standards of best practice for statistics in general, and for learning assessment methods and data in particular. The statements of general statistical principles/standards have been produced by agencies including the International Monetary Fund (IMF), the Organisation for Economic Cooperation and Development (OECD), UN Statistics Division, and the European Statistics System. The statements of principles/standards for learning assessment methods and data have been produced by agencies including the International Association for the Evaluation of Educational Achievement (IEA), the OECD, the American Educational Research Association (AERA), the American Psychological Association (APA), the National Council on Measurement in Education (NCME) and the International Testing Commission (ITC).

Further statements of standards of best practice in learning assessment can be found in the comprehensive technical documentation from all the major international assessments including the OECD's Programme for International Student Assessment (PISA) and Programme for International Assessment of Adult Competencies (PIAAC), and IEA's Trends in Mathematics and Science Study (TIMSS) and Programme for International Reading Literacy Study (PIRLS).

Though there is an abundance of existing material, the development of the ICP-LA is still warranted for a number of reasons. Firstly, the international code used to inform the monitoring of learning outcomes for SDG 4 should be independent and not associated with any one assessment or assessment agency. Secondly, there is a need for an international code that can accommodate the range of learning assessment activities that are undertaken throughout the world, and none of the existing articulations of standards can do this. Thirdly, the international code used for SDG monitoring should be situated within the UN framework for data quality. Finally, through a new code of practice (and associated supplementary material) UIS can emphasise its broader Education 2030 aims of building countries' capacity in learning assessments, linking learning assessment data into the broader network of education data, promoting more effective and efficient use of learning assessment data, and encouraging greater awareness of and literacy about learning assessment data in the general community.

Situating the ICP-LA in the broader SDG framework for data quality

Quality in a particular field can be articulated in a number of different ways. Quality principles are usually high-level and abstract statements about the ethics and expectations that underpin conduct in the field. Quality standards tend to be more concrete, describing how the principles are operationalised and serving as the basis for comparing or judging compliance. A code of practice is more practical and specific again, comprising guidelines to show how activities should be undertaken such that they adhere to standards and are in line with principles.

Any one articulation of quality must be aligned with the broader framework in which it sits. In the case of the ICP-LA, the broader framework is provided by the statements about the quality of statistics emanating from UN agencies, in particular the recently updated *UN Fundamental Principles for Official Statistics* and their associated implementation guidelines (United Nations General Assembly resolution 68/261; United Nations Statistics Division, 2015).

The fundamental principles are reproduced in Table 1 below. The titles of the principles (ie the text in bold) come from the implementation guidelines.

Note that since the UN principles are for official statistics, they are directed to agencies within or associated with national governments, whereas the principles on which the ICP-LA is based need to be equally relevant to all producers of data on learning outcomes.

Note also that the UN principles are quite general, since they are intended to guide producers of all kinds of official statistics, whereas the principles on which the ICP-LA is based can have a higher degree of specificity, since they are referring only to data collected via learning assessments.

These two points suggest that the UN's fundamental principles cannot be adopted as is for the ICP-LA. This does not mean, however, that they cannot provide the framework of principles for the ICP-LA.

Issue for further discussion

The GAML network should consider options for how the UN's principles can be adapted to provide an appropriate framework for the ICP-LA.

One possible option is to carry forward the titles for the principles (as given in Table 1), since they helpfully highlight key concepts. Text suitable for the ICP-LA context could then be formulated for each title.

Table 1: UN's fundamental principles of official statistics

1) Relevance, impartiality, equal access

Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

2) Professional standards, scientific principles, and professional ethics

To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

3) Accountability and transparency

To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

4) Prevention of misuse

The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

5) Sources of official statistics

Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

6) Confidentiality

Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

7) Legislation

The laws, regulations and measures under which the statistical systems operate are to be made public.

8) National coordination

Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

9) Use of international standards

The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

10) International cooperation

Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

Structure and content of the ICP-LA

The first section of the ICP-LA should give the introductory background information.

The second section of the ICP-LA should situate the code in the broader SDG framework of data quality. This section should include a mapping or linking of the principles that underpin the ICP-LA to the UN's *Fundamental Principles of Official Statistics* (see above discussion).[3]

The third section of the ICP-LA should cover broad concepts of quality in learning assessment, such as validity, reliability, fairness and fitness for purpose. This section should include definitions of these broad concepts and some initial discussion of the areas of activity in a learning assessment in which they play a part.

Issue for further discussion

There are a number of different ways to frame a discussion of key cross-cutting concepts in learning assessment. For example, the *PISA Technical Standards* refer to the overarching goals of consistency, precision, generalisability and timeliness (Organisation for Economic Cooperation and Development, 2015), and the *Standards for Educational and Psychological Testing* refer to the foundations of validity, reliability/precision and fairness in testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards in Educational and Psychological Testing, 2014).

The GAML network should consider the most appropriate way to frame a discussion of key cross-cutting concepts in the ICP-LA.

Issue for further discussion

The discussion of the concept of fitness for purpose – that is, the idea that the ultimate goal of a learning assessment is to produce data *that are appropriate for their designated roles or purposes* – will be an important part of the third section. This discussion will need to address the delicate balance between, on the one hand, achieving acceptable levels of technical rigour and, on the other hand, dealing with the realities of the financial and human resource limitations in which many learning assessments are undertaken. It will need to make the point that technical rigour cannot be sacrificed if it leads to assessment results that are systematically misleading or discriminatory. Compromise, however, is warranted where some threshold level of assessment and reporting quality is met – particularly in countries where there is limited available data or a significant need for reporting to support intervention and development.

[3] If, during the period in which the ICP-LA is being developed, other statements about or guidelines for data quality generally come out of UN agencies, these should be referenced in this section as well.

The GAML network is well placed to have input into this important discussion, because it includes representatives from the range of different assessment activities that are undertaken around the world.

The fourth section of the ICP-LA should cover practices in the following areas of activity:

- formulating the policy or research issues/goals that the learning assessment is intended to address;
- establishing and managing the institutional/organisational environment in which the learning assessment is designed;
- developing and implementing the technical and practical methods that yield, archive and secure the learning assessment data;
- preparing the learning assessment data products;
- disseminating and promoting the use of the data products.

This fourth section will be the operational core of the ICP-LA. Within each sub-section, some key points about standards of best practice should be made. Whenever a standard is provided, it should always be given with rationale that links back to the principles on which the ICP-LA is based and the cross-cutting concepts introduced earlier.

The four main sections of the ICP-LA described above will employ the extensive vocabulary of the theory and practice of learning assessments. This vocabulary should be explained in a glossary. In an electronic version of the ICP-LA, instances of the vocabulary should be hyperlinked to the relevant glossary definitions. In addition to including a complete glossary at the end of the ICP-LA, it may be worth considering having smaller, activity-specific glossaries dispersed throughout the operational information in the fourth section.

An outline of the proposed structure and content of the ICP-LA is given in Table 2 below.

Table 2: ICP-LA Structure and content outline

Front matter

Section 1: Background

Section discusses:

- SDG 4 context for measuring learning outcomes
- need for an ICP-LA

Section 2: Quality context for ICP-LA

Section discusses ICP-LA in relation to UN's *Fundamental Principles of Official Statistics*.

If, during the period in which the ICP-LA is being developed, other statements about or guidelines for data quality generally come out of UN agencies, these should be referenced in this section as well.

Section 3: Key quality concepts in learning assessment

Section discusses:

- reliability
- validity
- fairness
- fitness for purpose

Final structure of this section to be determined after discussion within GAM network.

Section 4: The international code of practice for learning assessment

Section discusses best practice in:

- using policy or research questions/goals to inform development of learning assessment
- establishing and managing the team responsible for designing and overseeing learning assessment
- learning assessment methods, including:
  - preparing and assessment framework
  - developing assessment and questionnaire items
  - translating assessment and questionnaire items
  - designing the test
  - sampling
  - field operations
  - data management
  - data analysis
  - reporting and dissemination

Supporting material

Including glossary

Lists of topics for each sub-section of *Section 4: The international code of practice in learning assessment* are given in Table 3 below. Note that these lists are not intended to be exhaustive.

Table 3: List of indicative topics and notes for sub-sections in Section 4: The international code of practice in learning assessment

Section 4: The international code of practice for learning assessment

Using policy or research questions/goals to inform development of learning assessment

Topics include:

- formulating policy or research question/goal that learning assessment can address, including involving all relevant stakeholders
- considering aspects of learning assessment design and implementation that are informed by policy or research question/goal

Establishing and managing the team responsible for designing and overseeing learning assessment

Topics include:

- key responsibilities of project team
- ensuring adequate financial and human resources
- ensuring adequate systems and infrastructure

Learning assessment methods:

Preparing an assessment framework

Topics include:

- the importance of developing an assessment framework as a tool that guides item and test development and provides a common language for discussing the assessment
- the scope and content of an assessment framework

Developing assessment and questionnaire items

Topics include:

- recruiting and training item developers – particularly important for assessment items
- using the assessment framework as a guide for item development
- pre-testing approaches for assuring item quality

Translating assessment and questionnaire items

Topics include:

- recruiting and training translators
- approaches to ensure cross-language comparability
- preparing and using translation guidelines

Test design

Topics include:

- developing a test design that ensures i) efficiency in sample sizes; ii) balanced test content; iii) stable measures over time

Sampling

Topics include:

- obtaining a complete and up-to-date sampling frame
- stratifying population represented in sampling frame
- using appropriate scientific sampling methods
- weighting
- assuring the quality of sampling and sample outcomes

Field operations

Topics include:

- recruiting and training test administrators
- preparing operations manuals
- administering the test
- maintain security of test material
- assuring the quality of test administration

Data management

Topics include:

- data modelling, i.e. representing relationships amongst objects within the sampling frame (e.g. relational data)
- data capture, i.e. the process whereby raw data are transferred to an initial database
- data cleaning, i.e. the process whereby the initial database is checked for errors, discrepancies and outliers
- data storage, security and life cycle

Data analysis

Topics include:

- methods for scaling assessment data
- applying sampling weights
- estimating uncertainty
- conducting special analyses to explore connections between assessment results and background characteristics and concepts of causality
- reproducible research methods

Reporting and dissemination

Topics include:

- developing a communication strategy that targets all interest groups
- preparing specific reports for different interest groups
- promoting appropriate and effective use of results by the different interest groups

The discussion in the sub-sections of *Section 4* will need to make clear statements about best practice while acknowledging and accommodating the variety of learning assessment activities undertaken throughout the world. For example, the discussion will need to acknowledge and accommodate:

- the range of purposes for learning assessments, from large-scale system monitoring to smaller-scale program evaluation;
- the variation in size, expertise and resourcing of project teams, including the fact that in some instances project teams may be supplemented by volunteers and in other instances particular activities such as sampling may be outsourced;
- the range of data capture process that are used, from manual data entry to scanning to online data capture;
- the variation in the time and expertise that project teams can commit to data cleaning;
- the different approaches to data analysis that are influenced by time, human and financial resources, and the policy or research questions/goals that the assessment is aiming to address.

Issue for further discussion

Developing an ICP-LA that covers everything in Table 2 and Table 3 in time for the first stage of SDG 4 monitoring activities may not feasible. In the first instance, developing an ICP-LA that is limited in scope but capable of meeting the immediate monitoring requirements of the UIS-led process should be considered.

Identifying what this limited scope might look like should be a discussion for all or part of the GAML network.

It will be important that the ICP-LA of limited scope presents as a cohesive whole yet can still accommodate the inclusion of additional content at a later date.

Style of the ICP-LA

The style of the ICP-LA should be influenced by the fact that it will need to serve the broad and varied range of individuals and groups with an interest in learning assessments. While specialised vocabulary and technical language will need to be used, they should not be used in such a way that the material becomes inaccessible to those who are approaching the field with less experience.

Supporting the interpretation and application of the ICP-LA

Beyond its role in informing the activities in the Data Quality Assurance Framework (as shown in Figure 1 above), it is expected that the ICP-LA will be a valuable general reference for groups and individuals working in the field of learning assessment.

Its utility as a general reference will be greater if it is supported by adequate discussion of the actions and contexts that guarantee or facilitate the application of best practice. If this discussion is provided, individuals and groups working in the field of learning assessment will be able to not only identify areas where their assessments may not be aligned with best practice, but also to understand the steps that need to be taken to move closer to alignment with best practice in the future.

Its utility as a general reference may also be greater if it is supported by discussion of approaches that are commonly adopted but that are not aligned with best practice, and the risks and consequences when these approaches are adopted. This suggestion is made based on the view that in some instances the most effective ways to warn against adopting an incorrect approach is to describe it and explicitly label it as incorrect.

Issue for further discussion

The discussions suggested above could be included in the ICP-LA itself, or in material that is supplementary to but associated with the ICP-LA.

The GAML network should consider the following options:

1) develop an ICP-LA that incorporates all supplementary discussions such as those suggested above;

2) develop an ICP-LA that is leaner and more focussed, and include discussions that support the interpretation and application of the ICP-LA, such as those suggested above, in separate supplementary documentation.

Section 3: The Evaluation of Alignment Process

This section introduces the Evaluation of Alignment Process (EAP) and proposes an approach to undertaking the EAP.

Alignment within the Evaluation of Alignment Process

Alignment[4], within the EAP: (1) confirms that there is significant conceptual overlap between the assessment program of the participating country and the UIS-RS, and (2) anchors the continuum of development represented by the assessment program to the UIS-RS. This approach supports inferences about educational progress on a common continuum without requiring participating countries to use a single assessment program (or a limited set of linked assessment programs) or curriculum approach.

Proposed Evaluation of Alignment Process

The EAP involves three stages. The first stage, referred to as *domain alignment*, is a review of top-level documents (e.g., assessment frameworks, blueprint, curriculum) to confirm that the assessment has conceptual overlap with the relevant UIS-RS domain. The second stage, referred to as *strand alignment*, collects evidence of the breadth and type of tasks that make up the assessment program via a task audit and uses this evidence to identify the extent to which the assessment program aligns with each of the strands within the UIS-RS domain. The third stage, referred to as *level alignment*, combines the evidence from the strand alignment stage to describe the coverage of the assessment program's continuum (e.g., from low to high achievement) relative to the UIS-RS. Each stage is described in the following sections.

The stages that make up the EAP are not for the purpose of excluding assessment programs or imposing a particular or narrow model of assessment. Rather, the EAP is a process of collecting evidence in a systematic and reliable way in order to locate a particular assessment on the UIS-RS. Particular focus is given to locating where the assessment program overlaps with the UIS-RS benchmarks that represent the SDG 4, indicator 4.1.1 minimum proficiencies defined at grades 2/3, the end of primary, and the end of lower secondary.

[4] *Alignment* is not the same as the psychometric method of *linking*: while one part of the alignment process can be a linking study (see the *level alignment* section) it is not the only way to reliably align an assessment program to the UIS-RS. The approach to alignment described here allows broader inclusion of assessment programs, though at potentially lower precision.

Issue for further discussion

Clear and internationally useful definitions of grades 2/3, the end of primary, and the end of lower secondary are required to ensure the EAP process is strongly aligned with reporting against SDG 4 indicator 4.1.1.

What are the definitions of each of these stages of schooling, and what are the relevant benchmarks on the UIS-RS for each?

Domain alignment

Each UIS-RS represents a domain: for example, reading. In each participating country, it is likely that the broad area of academic achievement being considered within an assessment program will extend beyond the scope of the UIS-RS. An assessment program that captures reading achievement, for example, may also capture writing, spelling, grammar, or punctuation achievement. Further, the structure of an assessment program that captures a wide range of the development within the domain is likely to change over that progression from low to high development: the assessment will prioritise different knowledge, skills, and understandings at different levels or for students participating at different ages. Domain alignment is, therefore, a stage to collect evidence about how the domain of interest is defined and operationalised within the participating country's assessment program. Top-level documents are reviewed to confirm that the assessment program being considered has overlap with the key elements of the UIS-RS domain and to define who is participating in the program.

Key documents of interest at this stage include curriculum documents, assessment, frameworks and blueprints, or other documents relevant within the participating country. The key information to be extracted is:

- How is the domain defined within the assessment program?
- What elements knowledge, skills and understandings that are assessed by the assessment program?
- What is/are the level/s of proficiency or capability the assessment program is intended to report on?
- Who participates in the assessment program (eg, what age and in what settings)?

In addition to documents sourced within the participating country, another source of information that feeds into this stage is the Catalogue of Learning Assessments (CLA). The CLA is a questionnaire based repository that provides internationally comparable information about learning assessments used within countries. The CLA contains information about how countries: (1) assess student capabilities using different learning assessments, (2) collect and analyse the data, (3) use results for policymaking, and (4) use results and reporting to understand and address the needs of the most vulnerable children. The most recent version of the CLA was explicitly designed to feed into the DQAF.

Issue for further discussion

Where the assessment program covers more than one distinct age or period of development, it may be necessary to undertake the following stages of the EAP for each age or period of development. This is because the mix of tasks used, strands included, and location on the UIS-RS will vary for each age/period of development. For example, a reading assessment program targeted at early primary school students may focus mostly on decoding skills, while at later ages; the same assessment program might be predominantly interpreting and reflecting tasks.

What criterion or threshold should be established that would require the strand alignment and level alignment stages to be conducted for each level of ability or age of student in the assessment program?

Should the strand alignment and level alignment stages be conducted separately for each reporting point of SDG 4.1.1 (grades 2/3; the end of primary; and the end of lower secondary) where the participating country reports that its assessment program is designed for children at one or more of those levels?

Strand alignment

In the UIS-RS, the constituent processes of the domain are referred to as *strands*. The UIS-RS provides a theoretically driven description of the kind of items that indicate the knowledge, skills, and understandings within each strand that make up the overarching UIS-RS capability (e.g., reading or mathematics). Each of these strands can be through of as lower-order factor that contributes to the overarching UIS-RS domain.

In the strand alignment stage, the assessment program being considered in the EAP is reviewed to describe the extent to which the strands are represented. The strand alignment stage collates evidence of the assessment program's coverage of the strands by first estimating the proportion of the assessment program devoted to each strand. This process begins with a *task audit* that provides evidence of the kinds of texts and items that represent the strands in the assessment program. From this, examples of task difficulty are evidenced at low, middle, and high difficulties on the assessment for each strand. For example, within an assessment program that is conducted with students within a single age or school-grade cohort, an example of the range of difficulties is collated. Where multiple ages or grade-cohorts are included (see issue for further discussion above), it may be necessary to provide examples for each age or grade-cohort. Where items are not available, it is plausible that descriptions of the assessment program's scale or proficiency levels could also be used.

Task audit

Tasks represent the breadth and type of *texts* and *items* used in the assessment program. This stage involves the participating country categorising the texts and items used in the assessment program according to the taxonomy provided by the UIS-RS. In the previous stage, evidence about the operationalisation of the domain within the assessment program was collected.

Where the assessment program covers very distinct ages or stages of development, it may be necessary to undertake the task audit for each age or stage as it is likely that the mix of tasks used will be very different in each.

Items within an assessment programs are likely to overlap across strands. This stage is about identifying the main strand that items are designed to assess. That is, each item will be described as primarily assessing one strand. Assessment frameworks or blueprints may also provide useful information about the primary purpose of items or the targeting of specific strands.

Utilising a range of tasks is an important way that an assessment program captures a broad range of student abilities (e.g., by including progressively more complex texts and items) and the full range of the domain as operationalised in the participating country (e.g., by using items and texts valued and appropriate to indicate the domain in the participating country). Building an understanding of this breadth within the assessment program is important for beginning the process of alignment.

At this stage of the EAP, the inclusion or omission of certain tasks does not preclude the alignment of the assessment program on the relevant UIS-RS. This stage of the EAP is used to create a robust understanding of the scope of the assessment program and to begin to build a shared understanding of the way the assessment program aligns with the UIS-RS. Building a strong understanding of the tasks included, for example, will help the participating country identify the range of difficulties assessed within each strand. The task audit phase collates evidence of the coverage of tasks by first estimating the proportion of the assessment program devoted to each type of defined task. Then example text and items (or descriptions of proficiencies or curriculum) are collected are collated for each task type.

The participating country will undertake the task audit and provide the information to the facilitating organisation. Several iterations may be necessary to build consensus that the task audit is representative of the assessment program relative to the UIS-RS. The facilitating organisation should at this stage be able to make general statements about the coverage of the domain by the assessment program.

Issue for further discussion

Some assessment programs may contain relatively few items to assess the relevant UIS-RS domain or the assessment program itself may be very short. Fewer items may impact the reliability of the estimate of proficiency.

Should there be a threshold number of items established before the assessment program is aligned? This would be in addition to the strand alignment task of estimating the proportion of items dedicated to each strand within the assessment program.

Task difficulty

The absence of items assessing a given strand does not preclude an assessment program from being aligned with the UIS-RS. At higher proficiency levels, however, the absence of items assessing a given strand may limit the assessment program from being aligned with higher levels of the UIS-RS.

The participating country will undertake the second part of strand alignment by using the evidence collected in the task audit. This is a necessarily collaborative process where the participating country works with the facilitating organisation to identify items representative of each strand and the full range of difficulty within each strand in the assessment program.

It is necessary during the strand alignment phase to build a strong shared understanding of the assessment program so that the final phase, level alignment, represents an accurate alignment between the assessment program and UIS-RS. At the end of this stage, the facilitating organisation will be able to make specific statements about the breadth of coverage of the domain, referring to the strands covered and to the range of item difficulties within each strand.

Level alignment

Level alignment is the anchoring of the participating country’s assessment program’s *continuum* on the UIS-RS. This process will locate the concept of low through high on the assessment program on some level or range of levels of the UIS-RS. It is likely that not all participating countries’ idea of low, medium, or high achievement will be the same, and any given country’s idea of low, medium, or high achievement will not necessarily reflect low, medium, or high achievement on the UIS-RS.

The simplest level alignment is conducted where an assessment has been (or will be) empirically linked to another assessment already aligned with the UIS-RS, or with the UIS-RS itself. Such an approach would provide sufficient information to align the proposed assessment with the UIS-RS.

In other cases, the level alignment process is dependent on reviewing information about the assessment program provided by the participating country in the earlier stages of the EAP It is likely that the kinds of information available will vary between countries and therefore two complementary approaches to level alignment are proposed and would be used together in different proportions, depending on the information available.

The facilitating organisation undertakes this stage using the information provided in the first two stages. This may be an iterative process where additional example items or supplemental information are requested from the participating country to assist with decision making.

The supplementary information may relate to statistical data (e.g., reporting on the proportion of students getting an item correct, or psychometric studies), or to assessment frameworks, blueprints, proficiency, or curriculum documents. At the end of this stage the facilitating organisation can make specific statements about the alignment of the assessment program with the UIS-RS.

Item mapping

Item mapping involves mapping items in the assessment program to the illustrative items that accompany the UIS-RS. This involves a qualitative rating of the difficulty of the items relative to the UIS-RS illustrative items. To complete this kind of level alignment, representative items that have been collected in the first two stages of the EAP are used. The supporting information provided by the participating country is used to locate a level on the UIS-RS where the items fit. Where the location is not obvious, the decision will be supported by supplementary information including:

- information about the age and/or grade level of the students being assessed
- descriptions of proficiencies or curriculum
- statistical information (e.g., the proportion of students who get that item correct)

Scale mapping

Scale mapping uses the descriptions of each achievement level of an assessment program (e.g., proficiency or benchmarks) to align them with the levels of UIS-RS. This kind of level alignment depends on the assessment program being able to yield a scale or continuum that describes the knowledge, skills, and understandings that are observed at certain points. This may range from standards in curriculum documents through to statistical information (for example minimum or proficiency standards yielded from psychometric studies or reporting).

Method

This section illustrates the proposed EAP and a summary of the proposed method is given in Figure 2. The UIS-RS for the domain of reading is used in all the examples that follow. The process for the mathematics domain will be similar but tailored to the domain (e.g., less focus on texts).

Issue for further discussion

The EAP will be a partnership between a participating country and a facilitating organisation. Who will be responsible as the facilitating organisation?

The EAP should be a transparent process with appropriate governance arrangements. Who will have oversight of the EAP?

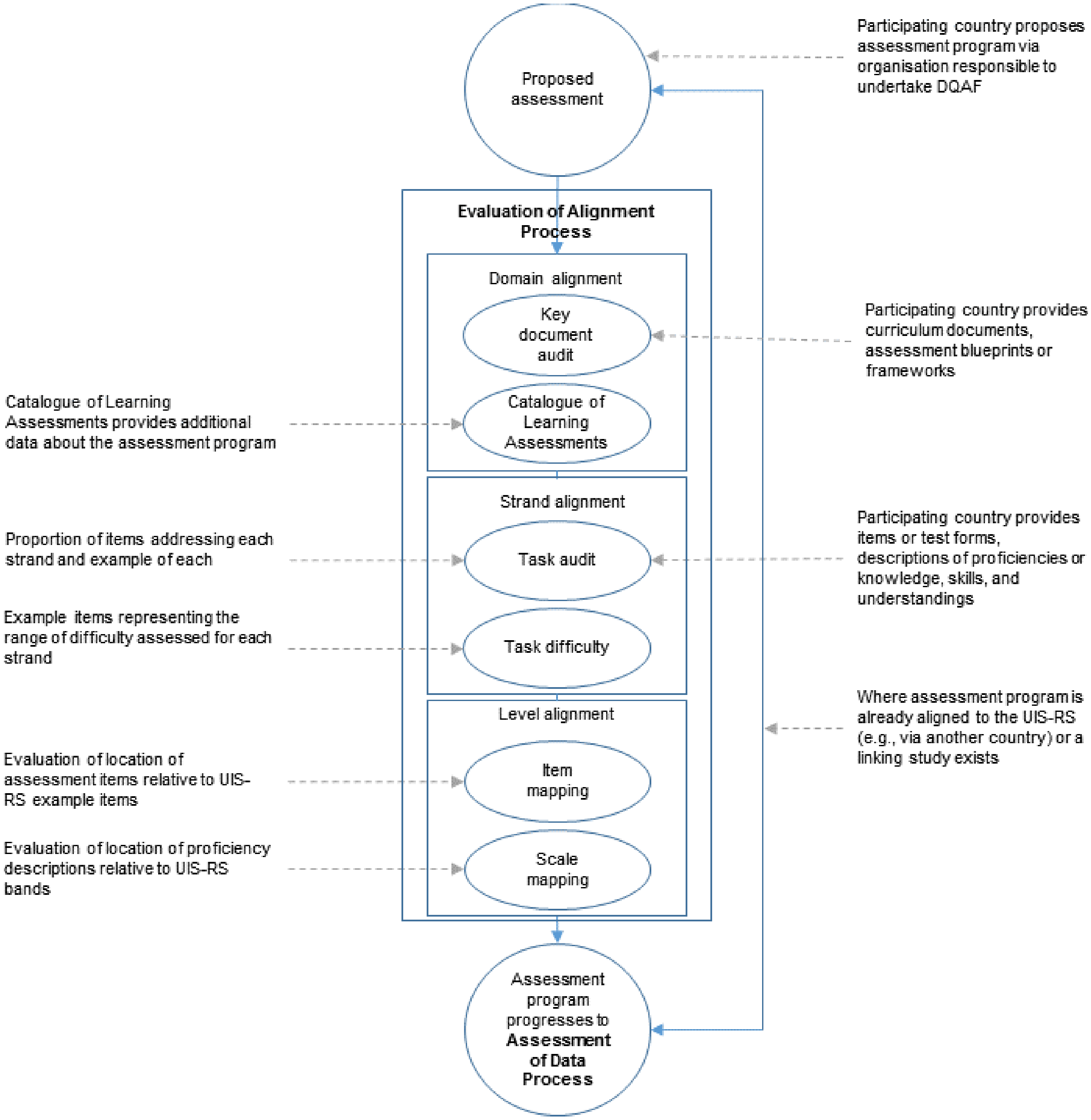

Proposed assessment
Participating country proposes assessment program via organisation responsible to undertake DQAF
Evaluation of Alignment Process
Domain alignment
Key document audit
Participating country provides curriculum documents, assessment blueprints or frameworks
Catalogue of Learning Assessments
Catalogue of Learning Assessments provides additional data about the assessment program
Strand alignment
Task audit
Proportion of items addressing each strand and example of each
Participating country provides items or test forms, descriptions of proficiencies or knowledge, skills, and understandings
Task difficulty
Example items representing the range of difficulty assessed for each strand
Level alignment
Where assessment program is already aligned to the UIS-RS (e.g., via another country) or a linking study exists
Item mapping
Evaluation of location of assessment items relative to UIS-RS example items
Scale mapping
Evaluation of location of proficiency descriptions relative to UIS-RS bands
Assessment program progresses to Assessment of Data Process

Figure 2: Summary of the EAP method

Domain alignment in the UIS-RS reading domain

Participating countries enter the EAP process having already completed the Catalogue of Learning Assessments (CLA). Where a participating country has not yet completed the CLA, this is the first step of the EAP. The most recent version of the CLA was explicitly designed to feed into the DQAF and is a robust entry point for a country that wishes to place an assessment program on the UIS-RS for the purpose of reporting against SDG 4 targets related to learning outcomes.

The participating country nominates the assessment program it wishes to align with the UIS-RS and this is assessed to confirm that the assessment program is the optimal nomination amongst all other assessment programs conducted by the country (e.g., are there other measures that are more appropriate or fit for purpose?)

Through the CLA, the breadth of learning assessments being conducted has been identified and sufficient information will be available to know when and how the assessment program was or is conducted, what is covered (e.g., the competencies assessed), and who participates.

The participating country also provides additional evidence, if not in the CLA, of:

- The definition of reading within the assessment program relative to the UIS-RS definition: "(reading) begins from the initial realisation that text contains meaning, passes through the capacity to interpret short written texts presenting familiar ideas, and moves on to the capacity to interpret and critically reflect on written texts with layers of subtle meaning that present unfamiliar ideas and draw upon a wide vocabulary" [5].
- The strands that make up reading within assessment program relative to the UIS-RS definition: "extract meaning from a text showing progress in decoding, retrieving information, interpreting information, and reflecting on the text".
- The levels of proficiency or capability the assessment program is intended to report on.
- A definition of who participates in the assessment program including whether it is a school based assessment program, or something else, and what grade or age of students/young people participate.

Strand alignment in the UIS-RS reading domain

The first stage of strand alignment is to conduct a task audit. The strands are:

- Decode text and understand concepts of print
- Retrieve information from text
- Interpret information in text
- Reflect on information in text

Task audit in the UIS-RS reading domain

Tasks are categorized according to the texts that students read and the items that students respond to in order to collect evidence of the knowledge, skills, and understandings that the assessment aims to test. This stage collates evidence of the assessment program's coverage of the strands by first estimating the proportion of the assessment program devoted to each strand. For example, what proportion of items assesses a student's ability to retrieve information from text?

[5] For further information, please see the Learning Progression Explorer.

For each strand, exemplar tasks in the reading domain are then categorised by the kind of text and items they represent. The kind of texts used help create a continuum of complexity and also represent the breadth of the reading domain. A text can be broken down according to *context*, *type,* and *format* and an item can have its own *format* (Council of Europe, 2001; OECD, 2016). The context reflects the purpose of the text or the audience it was written for and acknowledges that texts are not necessarily all written solely for the purpose of being used in an assessment program and can be categorised as:

- Personal - texts aimed at an individual's interests or at interpersonal communication (e.g., a section from a fictional novel).
- Educational - instructional texts focused on contributing to learning (e.g., a section from a text book).
- Occupational - texts focused on directing workplace processes or tasks (e.g., a job advertisement)
- Public -texts focused on conveying information broadly to society (e.g., information about public events).

The text type describes the predominant characteristics of the text and can be categorised as:

- Narrative -text that takes a subjective perspective to describe objects or action over time (e.g., plays)
- Exposition -texts that define or brings together phenomena to a complete whole (e.g., diagrams illustrating the parts of a whole system)
- Argument -persuasive or opinion texts describing the relationship among concepts or propositions (e.g., book review)
- Instruction -texts providing directions of what to do (eg, installation instructions)
- Transaction -texts that describe interpersonal communication coming to an agreed arrangement (e.g., emails exchanged to books a meeting date)

The text format is how the text is presented and is categorised as:

- Continuous -texts made up of sentences, paragraphs, pages, chapters, and books (or subsections) (e.g., essays)
- Non-continuous -texts made up words, images, graphics, and lists sometimes arranged as matrices (e.g., lists, or diagrams)

Item format is how the evidence of the student's ability is collected and can be categorized as:

- Not requiring expert judgement - (e.g., multiple choice or closed constructed response)
- Requiring expert judgement -(e.g., short or open constructed response)

The task audit phase collates evidence of the coverage of these tasks by first estimating the proportion of the assessment program devoted to each within each strand. For example, for the tasks related to the strand *Interpret information in text*, what proportion of texts is continuous and non-continuous? For each text and item type that is represented in the assessment program, example text and items, or descriptions of proficiencies or curriculum are collected.

Task difficulty in the UIS-RS reading domain

The second stage of the strand alignment is to ensure that the exemplar items collected represent the full range of difficult of the assessment program. For each strand that is represented in the assessment program, example tasks are provided by the participating country to illustrate low, middle, and high difficulties. This step provides the evidence to be used in the level alignment phase.

Issue for further discussion

This stage will not preclude an assessment program from being aligned with the UIS-RS. Where an assessment program insufficiently covers a strand, however, it might not be able to be aligned with higher levels of the UIS-RS.

Should some explicit standards be set? For example, from level 8 of the UIS-RS for reading interpreting and reflecting become significant strands with decoding being far less important. An example at level 8 may be that at more than 50 per cent of the items of an assessment program, at that level, should focus primarily in interpreting and reflecting.

Should a standard be set at each reporting point for SDG 4.1 (grades 2/3; the end of primary; and the end of lower secondary)?

Level alignment in the UIS-RS reading domain

Given strand alignment, the third phase of the EAP is level alignment. In the most simple case, where the participating country can show the assessment program is (or will be) linked[6] on the UIS-RS or to another assessment already aligned with the UIS-RS, the assessment program can immediately progress to being aligned. In other cases, the available items from the previous stages are used to locate anchor points on the UIS-RS. Where there is uncertainty, available items and descriptions from other, already aligned, assessment programs will be used to triangulate the anchor point.

[6] Linking refers to the formal psychometric approach to aligning the scaled within two assessment programs. Scales from different assessments measuring the same strand can be linked so that scores from one assessment can be expressed on the scale of another assessment. Parameter estimates can then be calculated and compared between countries (e.g., the proportion of students at a given level of ability) despite each using different assessment programs. Linking is rigorous but only a limited component in the EAP to ensure many countries can participate.

Item mapping in the UIS-RS reading domain

For each of the strands listed above, a qualitative assessment is made by the facilitating organisation of the evidence available for the assessment program covering levels on the UIS-RS. The example items available in the description of the UIS-RS are used to find commonalities that will anchor the assessment program to levels of the UIS-RS. Where there is underlining, additional items should be sought.

Scale mapping in the UIS-RS reading domain

Where there are limited or not items available, the process instead would rely on descriptions of the assessment program scale, other statistical data, or descriptions of assessment frameworks, proficiencies, or curriculum. In this case, the source of data to aid the facilitating organization in anchoring the scale or proficiency descriptions (for example) to the UIS-RS will be descriptions of *progression elements* from the UIS-RS. Progression elements are the description about *what is changing* as reading capability develops and are common across the strands. At each level of the UIS-RS, therefore, there is a description about the knowledge, skills, and understandings developing within each strand and these can be associated with descriptions available in the documentation supporting the assessment program.

Section 4: The Assessment of Data Process

This section introduces the Assessment of Data Process (ADP), then discusses: the different tools and methods in the ADP; how the ADP will be linked to the ICP-LA; how ADP decisions might be audited, adjudicated and reported; and the relationship between the ADP and other mechanisms for examining the quality of learning assessment systems and products.

Introduction to the ADP

As mentioned briefly in *Section 1*, the ADP is a process undertaken by the organisation with responsibility for ensuring the quality of data submitted by countries to UIS for reporting against the SDG 4 targets that refer to learning outcomes. It is expected that the ADP will be undertaken with the support of assessment contacts in the country, and will involve an examination of assessment documentation and databases, interviews with key informants, and perhaps observation of assessment activities, in order to determine the extent to which the methods and products from the country's implementation of the assessment are in line with best practices in learning assessment, as articulated in the ICP-LA.

Section 2 above outlines the proposed structure and content of the ICP-LA (see Table 2 and Table 3). The structure in Table 3 can inform the development of the ADP tools and methods, because guiding questions can be written for each heading to determine the scope of the information that the ADP will seek to discover.

The structure in Table 3 should also provide the overarching organising framework for ADP tools and methods. This will mean that the connection between the ADP and the ICP-LA is clear, which will reinforce the messages about everything that is required for best-practice learning assessment. It will also mean that should individuals or groups wish to use ADP tools and methods for their own evaluative purposes, they will be able to gain a better sense of the ADP context by referring to the ICP-LA. Some general characteristics of the ADP are presented with some discussion below:

- **The ADP will be reproducible and transparent.**

  Since the ADP will be a process that is undertaken many times, it will be vital that its tools and mechanisms can be applied consistently. It will also be vital that the ADP is transparent. This reproducibility and transparency will be achieved by making the development of the ADP consultative and collaborative, and by fully documenting ADP tools and processes and making the documentation publicly available. Ensuring the ADP is reproducible and transparent is important not only from an accountability perspective, but also because it will give others the best opportunity to make use proper use of ADP elements for their own evaluative purposes, should they wish to do so.

- **The ADP will be technically rigorous.**

  In *Section 2*, the discussion about how key-cross-cutting concepts will be addressed in the ICP-LA refers to the concept of fitness for purpose (see page 9 above). The ADP will need to operationalise this concept in its tools and methods. In other words, while the ADP needs to be technically rigorous, it should not be based on too narrow an understanding of what constitutes quality learning assessment data. If the ADP's criteria are too strict, it is likely that they will not be met by learning assessments from an unreasonable number of countries.

- **The ADP will be tailored to each country and the assessment program being used.**

  The ADP will consist of a number of different elements and approaches. It is likely that the same ADP approach will not be applicable in all contexts. It is also likely that in some instances, more than one possible ADP approach will be applicable. Moreover, in some instances the ADP may be a mere formality, because the quality of a country's assessment methods and products may have already been assured through another adequate process. For example, if a country wishes to submit its results from PISA for reporting against the UIS reporting scales, and its results were included in the PISA international report, then this country's methods and products must have satisfied the OECD's quality assurance measures, so they can be considered to satisfy the ADP as well.

  In general, the ADP approach adopted in each case should be the one that can obtain sufficient information about the quality of a country's learning assessment data in the most cost-efficient and timely manner, as this will help to guarantee the sustainability and suitability of the DQAF monitoring framework.

ADP tools and methods

The ADP will make use of the following tools:

- an evaluation framework and rubric
- checklists
- data collection instruments
- a template evaluation report
- a template improvement plan.

There will be range of methods through which these tools are used, including:

- desk-based review of databases from an assessment, as well as assessment instruments, results reports, operations manuals and any other technical documentation that is available

- remote questionnaire administration to key informants, including those involved in, for example, assessment design, item development, assessment administration, and data analysis
- in-country visits to conduct interviews and observations.

As mentioned above, it is expected that timeliness and cost-efficiency considerations will inform the decision about which tools/methods are used in each individual ADP implementation. The process for determining which combination of ADP tools and methods to use in each implementation should consider:

- The assessment for which the country wishes to submit results.
    - ➢ If the country wishes to submit results from an international or regional assessment of recognised technical quality, then the ADP may be able to be considered as satisfied.
    - ➢ If the country wishes to submit results from an assessment that has previously passed the ADP, then the current ADP may be able to be a 'light' version that seeks mainly to identify any key differences between previous assessment administrations and the current one.
- The availability of key documentation such as assessment framework or test blueprint, results report, technical report.
    - ➢ If the assessment does have some or all of this key documentation, then the ADP could be made up of desk-based review of the key documentation and the final database using the ADP checklists. Remote administrations of the ADP questionnaires to key informants could then fill any gaps in the information in the key documentation.
    - ➢ If the assessment doesn't have any of this key documentation, then the ADP could be made up of desk-based review of any available documentation using the ADP checklists, remote administration of ADP questionnaires to key informants, and, if required a site visit for interviews/observation.

Issue for further discussion

The GAML network and relevant external consultants should establish a process to determine which combination of ADP tools and methods will make each individual ADP implementation is as quick and cost-efficient as possible. It will be important that this process is documented and can be reliably applied, so that interested stakeholders understand why the ADP varies on a case-by-case basis.

Issue for further discussion

The GAML network and relevant external consultants should discuss and agree on the final set of tools and methods for the ADP.

In coming to this decision, it may be helpful for the GAML network and relevant external consultants to seek input from organisations such as the OECD, IEA, CONFEMEN, LLECE and SACMEQ about their quality monitoring and data adjudication tools and methods.

Auditing, adjudicating and reporting ADP decisions

While it is hoped that the ADP will be collaborative and transparent enough that a country will not be surprised by an ADP decision or think that it is unfounded, there nevertheless needs to be a mechanism for auditing ADP decisions and, in the case where a country disagrees with an ADP decision, a mechanism for adjudicating between the country and the organisation responsible for implementing the ADP.

The auditing and adjudication processes will need to be developed and undertaken by an independent and appropriately mandated group that has adequate expertise in the field of learning assessment. Documentation about these processes should be publicly available.

Issue for further discussion

The GAML network should consider how to establish governance protocols that include procedures to audit decision making and to resolve disputes or adjudicate decisions made within the ADP.

It may be worth considering establishing a panel or board to oversee, and working parties to undertake auditing and adjudication.

Countries should be encouraged to make at least some of the ADP findings public. Encouraging countries to make ADP findings available for the public will help to make the ADP more transparent overall. In addition, if ADP findings are made public, this may create an environment in which countries engage with each other about assessment activities, and thus benefit from peer-to-peer learning.

If countries are sensitive about ADP findings and prefer not to make them public, they should be at least encouraged to permit the circulation of findings within an agreed group from the GAML network, because this will enable areas of shared need in terms of capacity building to be identified and addressed in a cost-effective way. For example, if the ADP reveals that three countries in the same geographical region all need support to improve the analysis of their learning assessment data, then a workshop could be run for all three of them.

Issue for further discussion

The OECD's PISA for Development and the World Bank's Systems Approach for Better Education Results (SABER) are two initiatives that have made public the findings of their analyses of country capacity in learning assessment. The GAML network should draw on the experiences of these two initiatives to inform its handling of the sometimes delicate matter of making ADP findings public.

It may also be worth drawing on the experiences of other initiatives that include a component of country evaluation, the findings of which are made public. Regardless of the focus of these other initiatives, they may have some information that is useful for the ADP thinking. One example of an initiative that falls in this category is the IMF's Reports on Observances of Standards and Codes (ROSCs – see International Monetary Fund (n.d.)).

The relationship between the ADP and other initiatives that explore the quality of learning assessment systems and products

In the first instance the aim of the ADP is to determine whether the methods and products from a particular implementation of an assessment are sufficiently in line with best practice such that results from the assessment can be reported against the UIS reporting scales. But it also has a broader aim of building countries' capacity in learning assessment, in that ADP feedback will be framed so that countries can understand the areas in which they need to improve and how they might go about achieving that improvement. This will particularly be the case when a country's assessment does not satisfy the ADP, because then the country will be provided with an improvement plan and supported to implement it.

Since the ADP has this broader aim, its relationship with the other existing and planned initiatives that are committed to strengthening the monitoring of learning around the world needs to be considered. In particular, the relationship between the ADP, the World Bank's SABER, and the proposed Module 3 of the new version of UIS's Catalogue of Learning Assessments (CLA) should be considered. The scope and depth of the ADP should be such that it does not double-up on any information collected by SABER or Module 3 of the CLA, but rather that it complements and enriches the information obtained from these other two initiatives.

Issue for further discussion

The GAML network should consider how to best ensure that it complements SABER and Module 3 of the CLA.

The GAML network should encourage the groups responsible for SABER, Module 3 of the CLA and the ADP to work together to define objectives, scope and depth so that overlap is minimised and together the three initiatives can give the most complete picture possible of countries' assessment systems and the most comprehensive information to support countries to improve those systems.

Section 5: Next steps

The previous sections raised the following specific issues about the ICP-LA, the EAP and the ADP that should be discussed by the GAML network or associated external consultants. These issues are found in the "issue for further discussion" text boxes, above.

ICP-LA specific issues:

- adapting the UN's *Fundamental Principles of Official Statistics* so they can provide an appropriate framework for the ICP-LA
- framing the discussion about key cross-cutting concepts in learning assessment
- addressing the concept of fitness for purpose in the ICP-LA to capture the balance between achieving acceptable levels of technical rigour and dealing with the realities of human and financial resource limitations in which many learning assessments operate
- agreeing on the content of an initial version of the ICP-LA that is limited in scope but can be ready for the first stage of the US-led monitoring of learning outcomes for SDG 4
- deciding whether discussion that supports the interpretation and application of the ICP-LA should be included in the ICP-LA itself or provided in separate supplementary material.

EAP specific issues:

- What are the definitions of each of these stages of schooling, and what are the relevant benchmarks on the UIS-RS for each?
- What criterion or threshold should be established that would require the strand alignment and level alignment stages to be conducted for each level of ability or age of student in the assessment program?
- Should the strand alignment and level alignment stages be conducted separately for each reporting point of SDG 4.1.1 (grades 2/3; the end of primary; and the end of lower secondary) where the participating country reports that its assessment program is designed for children at one or more of those levels?
- Should there be a threshold number of items established before the assessment program is aligned? This would be in addition to the strand alignment task of estimating the proportion of items dedicated to each strand within the assessment program.
- The EAP should be a transparent process with appropriate governance arrangements. Who will have oversight of the EAP?

ADP specific issues:

- establishing a process to determine which combination of all available ADP tools/methods is used in each specific ADP implementations
- agreeing on the final set of ADP tools and methods
- instigating processes for auditing and adjudicating ADP decisions
- handling the potentially delicate matter of making ADP findings public
- ensuring complementarity between the ADP, SABER and Module 3 of the CLA.

In addition to these specific issues, the GAML network should consider the following more general questions about the ICP-LA, the EAP and the ADP.

ICP-LA general issues:

- Who will develop the final outline for the ICP-LA?
- What kind of review, feedback and endorsement process will the final outline of the ICP-LA need to go through before work on the substance begins?
- Who will write the substance of the ICP-LA and any supplementary material?
- What kind of review, feedback and endorsement process will the substance of the ICP-LA and any supplementary material need to go through?

EAP general issues:

- Who will finalise the specification of the EAP method, including developing materials for the assessment of the three stages?
- Should a separate method document be drafted for the mathematics domain?
- Who will be responsible for implementing the EAP?

ADP general issues:

- Who will progress the ADP thinking?
- Who will develop the ADP tools and methods?
- Who will implement the ADP?

We suggest that these questions and issues should be addressed by a GAML Task Force established to support further work on the DQAF. Terms of Reference for such a task force have been provided in a separate meeting document.

Bibliography

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards in Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. Washington DC: AERA.

Australian Bureau of Statistics. The ABS Data Quality Framework. Retrieved 30 May, 2016, from http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Quality:+The+ABS+Data+Quality+Framework

Australian Bureau of Statistics, & National Statistical Service. Data Fitness: A guide to keeping your data in good shape.

Caldwell, N., Foy, P., Martin, M. O., Mullis, I. V. S., & Sibberns, H. (1999). Technical Standards for IEA Studies. M. O. Martin, K. Rust & R. J. Adams (Eds.), Retrieved from http://www.iea.nl/publication_list.html?&no_cache=1

Clarke, M. (2012). What Matters Most for Student Assessment Systems: A framework paper SABER Working Paper Series, Retrieved from http://saber.worldbank.org/index.cfm?indx=8&pd=5&sub=4

Council of Europe: Language Policy Unit. (2011). Common European Framework of Reference for Languages: Learning, teaching, assessment. Retrieved from http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

Educational Testing Service. (2014). ETS Standards for Quality and Fairness. Retrieved from https://www.ets.org/about/fairness/

Ehling, M., Körner, T., Bergdahl, M., Elvers, E., Földesi, E., Kron, A., *et al.* (2007). EuroStat Handbook on Data Quality Assessment Methods and Tools. Manfred Ehling & T. Körner (Eds.), Retrieved from http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf

European Statistical System. (2011). European Statistics Code of Practice. Retrieved from http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

European Statistical System. (2011). Quality Assurance Framework for the European Statistical System. Retrieved from http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

Gregory, K. D., & Martin, M. O. (2001). Technical Standards for IEA Studies: An Annotated Bibliography. Retrieved from http://www.iea.nl/publication_list.html?&no_cache=1

Independent Expert Advisory Group on a Data Revolution for Sustainable Development. (2014). A World that Counts: Mobilising the Data Revolution for Sustainable Development.

Inter-Agency and Expert Group on Sustainable Development Goal Indicators. (2016). Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators.

Inter-Agency and Expert Group on Sustainable Development Goal Indicators. (2016). Update on the work to finalize the proposals for the global indicators for the Sustainable Development Goals.

International Institute for Educational Planning. (2016). Financing Education 2030 - The IIEP Letter (Vol. 1): United Nations Educational, Scientific and Cultural Organisation.

International Monetary Fund. Data Quality Reference Site, from http://dsbb.imf.org/Pages/DQRS/DQAF.aspx

International Monetary Fund. (2006). Data Quality Assessment Framework: A Factsheet.

International Monetary Fund. (2012). Data Quality Assessment Framework - Generic Framework.

International Monetary Fund. (2013). The General Data Dissemination System - Factsheet. Washington DC: International Monetary Fund.

International Monetary Fund. (2013). The General Data Dissemination System : guide for participants and users.   Retrieved from http://dsbb.imf.org/pages/gdds/home.aspx

Izard, J. (2005). Module 6: Overview of test construction. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

Izard, J. (2005). Module 7: Trial testing and item analysis in test construction. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

Johansone, I. (2012). Operations and Quality Assurance. In Michael O Martin & Eva V.S. Mullis (Eds.), Methods and Procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP_Operations_Quality_Assurance.pdf.

Kenneth N. Ross, T. Neville Postlethwaite, Marlaine Lockheed, Aletta Grisay, & Gabriel Carceles Breis. (1990). Improving data collection, preparation and analysis procedures: a review of technical issues. In K. N. Ross & L. Mählck (Eds.), *Planning the Quality of Education: The collection and use of data for informed decision-making*. Paris & Oxford: UNESCO International Institute for Educational Planning & Pergamon Press.

Livingstone, I. D. (2005). Module 2: From educational policy issues to specific research questions and the basic elements of research design. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

National Statistical Service. Data Quality Online. Retrieved 30 May, 2016, from https://www.nss.gov.au/dataquality/index.jsp

Organisation for Economic Cooperation and Development. (2012). Quality Framework and Guidelines for OECD Statistical Activities.   Retrieved from http://www.oecd.org/std/qualityframeworkforoecdstatisticalactivities.htm

Organisation for Economic Cooperation and Development. (2013). Technical Report of the Survey of Adult Skills (PIAAC) - pre-publication copy: OECD.

Organisation for Economic Cooperation and Development. (2014). PIAAC Technical Standards and Guidelines.: OECD.

Organisation for Economic Cooperation and Development. (2014). PISA 2012 Technical Report Retrieved from https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

Organisation for Economic Cooperation and Development. (2015). *PISA 2015 Technical Standards.*: OECD.

Postlethwaite, T. N. (2005). Module 1: Educational research: some basic concepts and terminology. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

Rashtriya Madhyamik Shiksha Abhiyan Technical Cooperation Agency. (2016). Large-Scale Learning Assessments: A Handbook for the Indian Context. Retrieved from http://rmsaindia.org/en/?option=com_pdf&view=__recentdocs&catid=all&type=-1&main=0&Itemid=224

Ross, K. N. (2005). Module 3: Sample design for educational survey research. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

Schleicher, A., & Saito, M. (2005). Module 10: Data preparation and management. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

Siniscalco, M. T., & Auriat, N. (2005). Module 8: Questionnaire design. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

Statistics Canada. (2002). Statistics Canada's Quality Assurance Framework. Ottawa: Ministry of Industry.

UNESCO Institute for Statistics. (2016). Meet the Education 2030 Data: A guide to the indicators to monitor SDG 4 - Education 2030 (Brief about indicator 4.1.1) Retrieved 10 June, 2016, from http://www.uis.unesco.org/Education/Pages/meet-the-education-2030-data-indicator-4-1-1.aspx

UNESCO Institute for Statistics. (2016). Meet the Education 2030 Data: A guide to the indicators to monitor SDG 4 - Education 2030 (Brief about indicator 4.1.2) Retrieved 10 June, 2016, from http://www.uis.unesco.org/Education/Pages/meet-the-education-2030-data-indicator-4-1-2.aspx

UNESCO Institute for Statistics, & World Bank. (22 October 2012). Ed-DQAF wiki Retrieved 30 May, 2016, from http://dqaf.uis.unesco.org/index.php?title=Main_Page

United Nations General Assembly resolution 68/261. Fundamental Principles of Official Statistics. *A/RES/68/261 (from 29 January 2014)*.

United Nations Statistics Division. (2015). United Nations Fundamental Principles of Official Statistics: Implementation Guidelines. Retrieved from http://unstats.un.org/unsd/dnss/gp/impguide.aspx

Withers, G. (2005). Module 5: Item writing for tests and examinations. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

Wolf, R. M. (2005). Module 4: Judging educational research based on experiments and surveys. In K. N. Ross (Ed.), Quantitative Research Methods in Educational Planning. Paris, France: UNESCO International Institute of Educational Planning. Retrieved from http://www.iiep.unesco.org/en/library-resources/briefs-papers-tools.

World Bank. (n.d.). SABER - Student Assessment Questionnaires and Rubric: World Bank.