# Principles of Good Practice in Learning Assessment

United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
FOR
STATISTICS

ACER

# Table of contents

# List of abbreviations

| | |
|---|---|
| ACER | Australian Council for Educational Research |
| ACER-GEM | ACER Centre for Global Education Monitoring |
| CTT | Classical Test Theory |
| GAML | Global Alliance to Monitor Learning |
| GPE | Global Partnership for Education |
| GP-LA | Principles of Good Practice in Learning Assessment |
| IEA | International Association for the Evaluation of Educational Achievement |
| IRT | Item Response Theory |
| NCERT | National Council of Educational Research and Training |
| OECD | Organisation for Economic Co-operation and Development |
| PISA | Programme for International Student Assessment |
| SDG | Sustainable Development Goal |
| TIMSS | Trends in International Mathematics and Science Study |
| UIS | UNESCO Institute for Statistics |
| UN | United Nations |
| UNESCO | United Nations Educational, Scientific and Cultural Organisation |

# List of figures



# List of tables

# Acknowledgements

The 'Principles of Good Practice in Learning Assessment' document was authored by the Australian Council for Educational Research (ACER), Centre for Global Education Monitoring (ACER-GEM), in collaboration with the UNESCO Institute for Statistics (UIS), Global Alliance to Monitor Learning (GAML). Valuable input was provided by the GAML secretariat at UIS (lead by Silvia Montoya), and the GAML Assessment Implementation Task Force (chaired by Esther Care, Brookings Institution), and Task Force 4.1 (chaired by Marguerite Clarke, World Bank).

The principles of good practice in learning assessment described in this document draw on the extensive experience of ACER in planning, developing and conducting large-scale assessments. Most notably the OECD Programme for International Student Assessment (PISA), for which ACER was the leading consortium partner during five cycles of implementation over a period of more than 12 years. Sincere recognition is made of the work of the very many people whose efforts have contributed to the development, conduct and reporting of assessments over the years, on whose accumulated experience and expertise this present document is based.

Major contributions to this document were made by Stephanie Templeton, Prue Anderson, Alla Berezner, Jorge Fallas, Petra Lietz, Clare Ozolins, Jeaniene Spink, Claire Scoular, Naoko Tabata, Alvin Vista, Maurice Walker, and Lam Winson. Raymond Adams, Ursula Schwantner, Ross Turner and Charlotte Waters were responsible for the overall conceptualisation and development. Maurice Walker initiated the development of the 14 key areas as a key approach to a robust assessment program of ACER-GEM. The document also builds upon previous work undertaken for NCERT in India (NCERT, 2015), and the Public Education Evaluation Commission in the Kingdom of Saudi Arabia (PEEC & ACER, 2016).

# Introduction

In September 2015, the United Nations (UN) adopted the 2030 Agenda for Sustainable Development, otherwise known as the Sustainable Development Goals (SDGs). These goals are a declaration by the international community to take action in ending poverty, protecting the planet and building peace. There are 17 goals in total, with education the focus of Goal 4: "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (Inter-Agency and Expert Group on Sustainable Development Goal Indicators, 2016, p. 19).

Key to SDG 4 are the notions of quality in education and ensuring that all members of society have equitable access to education opportunities. Quality in education has been defined by SDG 4 in a number of ways (Inter-Agency and Expert Group on Sustainable Development Goal Indicators, 2016). This includes the provision of a safe learning environment for all (SDG target 4.a), and the availability of qualified teachers (SDG target 4.c). However, the major component of the definition of quality for SDG 4, is that various members of the population should attain a minimum level of knowledge and skills in certain learning domains. For example, grade 2/3 students with minimum proficiency in reading and mathematics, or youth/adults with information and communication technology (ICT) skills. Reporting progress towards these targets will thus require countries to assess learning outcomes.

An important mechanism for establishing and monitoring education quality at the system level is large-scale assessments. Large-scale assessments focus on defined learning domains (e.g. reading, mathematics). The content of a learning domain is based on commonly acknowledged theories of learning in the relevant domain (and typically defined in the assessment framework), with a focus on whether the learner can apply the skills and concepts that she or he has acquired and learned. The content of the assessed learning domain (assessment framework) may be referenced to a national curriculum. Large-scale assessments can be international, regional or national in scope. Large-scale assessments focus on a particular population (e.g. Grade 2 or Grade 3 students in schools, students aged 15, children aged 15), they can be sample-based or conducted as a census.

Large-scale assessments are conducted for a range of reasons, including:

- to establish and describe the knowledge and skills of a particular population (sample or census) in a learning domain;
- to monitor progress in learning outcomes over time or between grades;
- to investigate associations between achievement and contexts in which learning takes place;
- to quantify differences in learning outcomes between sub-populations (e.g. girls and boys);
- to report, in the case of census-based assessments, school or individual level results.

To be effective, large-scale assessments need to gather data that provide an accurate reflection of the present situation. As such, the management of data quality plays a central role in SDG 4.

At the national level, this involves the development of national strategies for large-scale assessments, education data and the commitment to building assessment and statistical capacity. At the international level, this involves a participatory approach to the development of international standards and methodologies, the provision of diagnostic tools and guidelines, and support in capacity development (UNESCO Institute for Statistics, 2016).

The Principles of Good Practice in Learning Assessment (GP-LA) are a central element of the international commitment to the management of SDG 4 data quality for learning outcomes. The GP-LA is an independent articulation of good practices that accommodates the diversity of large-scale learning assessment activities being undertaken throughout the world. Within SDG 4 reporting processes it serves two purposes:

- First, it serves as the conceptual framework to evaluate the quality of large scale assessments and data from these assessments submitted for SDG 4 reporting. By outlining key principles of assessment quality, it helps countries to achieve technical rigour while also allowing flexibility so that countries can set their own assessment priorities within national contexts.
- Second, the principles described in the GP-LA will support the diagnosis of country-level capacity to develop, implement and use data from large-scale assessments, and the formulation of capacity-development plans. In a case where large-scale assessment data submitted for SDG 4 reporting does not meet the standards required for reporting, the GP-LA will inform the development of an improvement plan and help target technical support.

Beyond its specific role within the SDG 4 monitoring framework, the GP-LA also serves as a general reference for any groups or individuals working within the field of learning assessment. The practices described in efficiently planning, developing and implementing a robust large-scale assessment program, with the aim to effectively using the data for education system monitoring and evidence-based education policy.

The GP-LA is a statement of principles, designed to be advisory for developing and implementing assessment programs. The statements are deliberately cast at a level that is general in nature so they are applicable in various large-scale assessment contexts and settings (e.g. with international, regional or national focus; school-based as well as household-based assessments). The document does not set standards, since these will be context dependent. Concrete standards for determining the successful application of the good practices outlined in the GP-LA will be operationalised and formulated for the purpose of evaluating data quality for SDG 4 monitoring as part of the 'Evaluation of Data Collection'. In addition, more practical 'How-to Guides' will be developed to illustrate the steps in implementing an assessment program, and to provide examples demonstrating the principles of good practice in action.

## Organisation of the GP-LA

The GP-LA contains three main sections:

1) The first section, *Broader quality Context for the GP-LA,* describes the UN Fundamental Principles of Official Statistics as a wider framework for the principles discussed in the GP-LA.
2) The second section, *Key Quality Concepts of Learning Assessment*, describes the cross-cutting quality concepts of fitness for purpose, clarity and consistency of purpose, objectivity and independence, transparency and accountability, technical rigour, ethicality and fairness.
3) The third section, *Good Practice in Learning Assessment*, describes the 14 key areas of a robust assessment program which can also be viewed as the 'lifecycle' of an assessment program. Each of the key areas contains:
   - a statement describing the objective and products of the area considering good practice in learning assessment;
   - a breakdown and further explanation of that statement that defines terms, and elaborates on characteristics and uses of the product described in the objective; and
   - a description of how to achieve the objective and products through good practices and processes.

The GP-LA aims to be clear and accessible to all users no matter how familiar they are with learning assessment. However, technical language and specialised vocabulary is sometimes necessary. A glossary of terms is provided in the Appendix to clarify meanings where needed.

# Broader quality context for the GP-LA: The UN Fundamental Principles of Official Statistics

As an independent articulation of good practice in learning assessment, and in support of the international commitment to the monitoring of learning outcomes under SDG 4, the GP-LA has been set within a broader quality framework, in particular the United Nations Fundamental Principles of Statistics (see Table 1).

The UN Fundamental Principles of Statistics define professional and scientific standards for the generation of official statistics across all aspects of governance. They aim to help nations improve and build their processes and capacity in planning, collecting, and disseminating statistics, in order to inform evidence-based decision making (United Nations Statistics Division, 2015).

The GP-LA has similar aims. However, there are two important differences between the GP-LA and the UN Fundamental Principles of Official Statistics. First, the Fundamental Principles are directed towards agencies within or associated with national governments, whereas the GP-LA is equally relevant for all bodies involved in collecting and disseminating data on learning

outcomes. Second, the Fundamental Principles are defined in a more general way, whereas the GP-LA defines standards specifically for learning assessment and issues involved in generating data on learning outcomes. As a result, the GP-LA uses the Fundamental Principles as a reference but not as a direct framework. Listed below are the Fundamental Principles. The relationship between the UN Fundamental Principles of Official Statistics and the content of the GP-LA is described in more detail in the Appendix.

**Table 1: UN Fundamental Principles of Official Statistics**

| |
|---|
| *Principle 1*: Fundamental principles 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. |
| *Principle 2*: 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. |
| *Principle 3*: 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. |
| *Principle 4:* The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics |
| *Principle 5*: 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 |
| *Principle 6*: 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. |
| *Principle 7:* The laws, regulations and measures under which the statistical systems operate are to be made public. |
| *Principle 8:* Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system. |
| *Principle 9:* 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. |
| *Principle 10:* Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries. |

*Source: (United Nations Statistics Division, 2015)*

# Key quality concepts for learning assessments

The following describes the key quality concepts for learning assessments, they are expressions of the UN Fundamental Principles of Official Statistics specifically in relation to large-scale learning assessments. While the UN Fundamental Principles of Official Statistics express broad concepts of quality that need to be relevant to all types of statistics, concepts of quality in learning assessment can be expressed in specific terms. This section of the GP-LA outlines the key quality concepts in learning assessment that cut across all areas of a robust assessment program. These concepts are not mutually exclusive, but interact and overlap to form a comprehensive definition of quality in learning assessment.

## Fitness for purpose

'Fitness for purpose' describes the concept that the ultimate goal of a learning assessment is to generate data *that are appropriate for their designated purposes*. The learning outcomes data that are used for SDG 4 monitoring/reporting are derived from various international, regional and national learning assessments, each having their own designated purposes. Hence, an added concern for reporting on SDG 4 learning outcomes is the comparability of the data across the various assessments and contexts (regions/countries). Thus, 'fitness for purpose' in this global context of SDG 4 monitoring, is concerned with the appropriateness of the data for the purpose of international reporting on learning outcomes. It aims to strike a balance between technical rigour and the practical implications of using and comparing data from a variety of existing learning assessments.

## Clarity and consistency of purpose

The purposes of large-scale assessments of learning can be multiple, for example establishing the knowledge and skills of a population, monitoring learning progress over time, monitoring equity, exercising accountability or targeting funding. An assessment may address one or more purposes. As such, clearly stating the purpose(s) of an assessment is essential. An outline of the reason for the collection of data and its intended uses will serve as a guide for all future decisions made in relation to the assessment program, as well as a standard against which to evaluate the overall effectiveness of an assessment.

All decisions made in relation to an assessment program should be consistent with its stated purpose. This will help maintain consistency across all areas of the assessment and ensure that the final results are relevant and useful for education policy makers and other stakeholders. Technical decisions include, for example, who to assess, what to assess, the format and design of the assessment, how it is implemented in the field, which analyses are used, and how results will be reported and disseminated. Operational decisions include staffing, timing, and the resources dedicated to an assessment.

The step from translating assessment findings into education policy or education system reforms should be based on evidence from the data gathered.

## Objectivity and independence

As the indicators for SDG 4 illustrate, quantitative data occupy a position of authority in the global landscape when describing phenomena like education quality and equity. Whilst this position is based on the idea that quantitative data provide an unbiased, value-free measure of these phenomena, data are not inherently objective; it can be manipulated at various stages of the assessment process so that the results tell a version of the story that supports certain interests.

In order for stakeholders to trust the results of an assessment, the collection of the data must be objective. Special interests should not influence how data is collected if it is to affect the

technical rigour and scientific standing of the learning assessment. Hiring learning assessment professionals and consulting experts can help to avoid inappropriate influence.

Similarly, the interpretation of the data should also be objective and independent. Interpretation should be in alignment with the purpose of the assessment and the data generated, and it should not go beyond the constraints placed upon it by the assessment's design. For example, excluding certain groups of the population from participation can have important consequences on the generalisability of the results.

## Transparency and accountability

All aspects of an assessment program should be open to outside scrutiny. This means that the assessment methodology, implementation processes and data analysis methods and procedures should be clearly described and publicly available. By justifying the decisions made in relation to the assessment methodology, implementation and analysis, the results are not only verifiable by other experts in the field, but they are more robust to criticism. This also helps contribute to the objectivity of the results.

Transparency means that an assessment program is held accountable to its stakeholders. It ensures that assessments are objective, feasible, timely, technically robust, consistent with their intended purposes, use resources efficiently, and are useful for education policy making. It also ensures that the assessment program adheres to laws governing the generation of education data and statistics.

## Technical rigour

It is essential that assessment methodology, analysis and interpretation of data follow scientific principles. The aim of all key areas should be technical rigour so that inferences drawn are valid and their level of certainty can be determined. That includes for example rigorous scientific sampling procedures, selection of appropriate analytic methods, well-constructed contextual instruments, and valid and reliable assessment tools. The principles described in this document are aimed at ensuring technical rigour in all phases of an assessment program in order for it to be technically robust.

## Ethicality and fairness

The broad goal of research ethics and fairness is to ensure that no harm is done to individuals or groups as a result of a research study, for example, as a result of an assessment of learning outcomes (American Educational Research Association, 2011). This broad concept of 'do no harm' should be considered in all areas of an assessment program, from defining the purposes and its development, to administration and data management, to analysis, reporting, interpretation and communication. Moreover, professional competence, integrity, and responsibility are considered important ethical concepts (American Educational Research Association, 2011), forming the basis of all the key quality concepts discussed in this document, i.e. clarity and consistency of purpose, objectivity and independence, transparency and accountability, and technical rigour.

For learning assessments ethics and fairness principles related to the *participant* seem most important. A primary principle in this regard is to ensure the *confidentiality* of the participants. This includes for example anonymising data for public release, the secure storage of test data whether they be in the form of completed paper-based tests or digital databases, and having staff and contractors sign confidentiality agreements.

Assessments should be designed and administered considering the *well-being* of participants. This includes, for example, considerations about the timing and length of the assessment, which should be appropriate for the target population. Such decisions require a balance between the scientific needs of the assessment design, and the testing load that participants can handle. Standardised test administration procedures and instructions should be developed in a way that helps to create a welcoming atmosphere for participants, so as to prevent test anxiety and to encourage those completing the assessment to participate and to perform at their best.

A principle of fairness that is strongly related to technical rigour is the aim to *minimise measurement bias.* Measurement bias is where a test and/or contextual instrument consistently discriminates against a particular group of participants for reasons unrelated to the learning domain being assessed or background data being collected. For example, complex language in test instructions may prevent some participants from understanding what they are required to do to answer the question. As such, these participants may be unable to fully demonstrate their skills and knowledge or personal background and attitudes.

Another important ethics and fairness principle for large-scale assessments that relates to the participant as well as to the quality concepts of clarity and consistency of purpose and technical rigour, is *inclusiveness.* Inclusiveness means to design assessments to be relevant for as many members of the target population as possible. Thus, the concept of inclusiveness has an impact on the definition of the target population (e.g. students in grades 2 and 3; all children between age 5 and 16) as well as on practical matters such as the translation of the assessment tools into different national/regional languages or accommodating special needs (e.g. providing large-print test and questionnaire forms for people with visual disabilities). Ideally the judgement about inclusiveness is made in consideration of the main purposes of the assessment, the precision of the data required to fulfil these purposes and the operational costs for collecting data of such precision, especially when resources are limited.

# Good practice in learning assessment

Good practice in learning assessment ensures that key quality concepts are met at all stages of an assessment program.

Figure 1 shows the 14 key areas of a robust assessment program, which can also be viewed as the 'lifecycle' of an assessment program. Key areas range from identifying and defining policy goals and education issues, to designing and implementing the assessment, through to analysing and reporting the data, to informing the initially defined policy goals and issues.

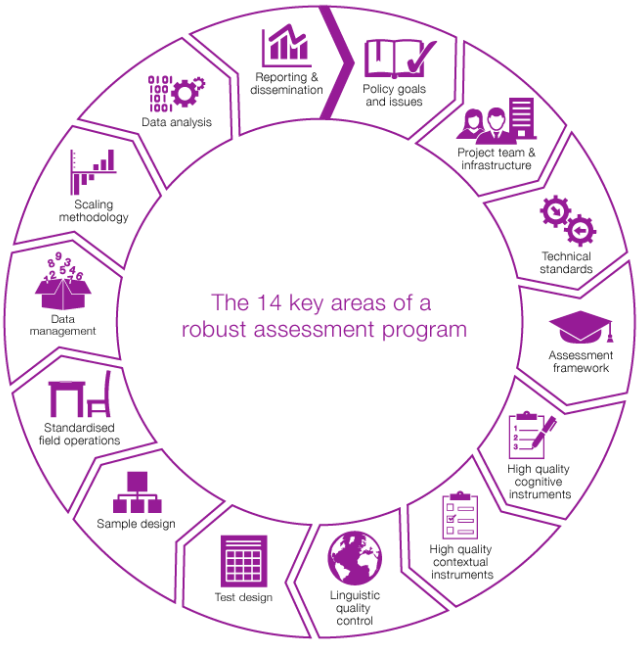


**Figure 1: The 14 key areas of a robust assessment program**

To be robust and effective, it is important that assessments use well-founded methodologies in each of these key areas. This will help assessment agencies to yield high quality data and, in using these data, to provide meaningful analysis that are of use for educational monitoring, policy and practice.

For each key area, first the *objective* is described, and then *how the objective can be achieved*.

## Key Area 1: Formulating policy goals and priorities to be addressed with the learning assessment

### Objective

| |
|---|
| Clearly articulated policy goals and measurement priorities that are relevant to key stakeholders and inform the content, design and scope of the assessment program. |

***Clearly articulated policy goals and measurement priorities…***

The term 'policy goals' refers to the overall purpose/s of the assessment. The term 'measurement priorities' refers to the specific statistical objectives addressed by the assessment. These goals and priorities should be formally stated as they will guide decisions about the content, design and scope of the assessment. They also enable officials and assessment agencies to communicate the assessment program to others (Johnson & Christensen, 2014). As such, the goal and priority statements need to be clearly articulated.

***…that are relevant to key stakeholders…***

Assessment results of high technical quality are only useful if they speak to stakeholders' education priorities. Key stakeholders may include: government representatives at the national and sub-national levels; planning advisors, policy developers and curriculum developers; representatives of school sectors, e.g. public or private institutions; teacher professional bodies and teacher training institutions; academics, donors, schools, parents, and children. No one assessment can cater to the needs of all stakeholders, therefore it should be clear which stakeholders will be the primary users of the assessment data. These stakeholders should then be consulted throughout the assessment program by way of a steering committee.

***…and can inform the content, design and scope of the assessment program.***

By articulating the policy goals and measurement priorities of the assessment, the steering committee and assessment agency are provided with a basis upon which they can make specific plans for the content, design and scope of the learning assessment. By aligning the content, design and scope of the assessment with the policy goals and measurement priorities, it is more likely that the assessment results will be considered in education policy and practice. The goal statements can also be used to evaluate the suitability of decisions made about the assessment.

## Achieving the objective

**Identify education priorities.** Policy goals and measurement priorities can be informed by identifying the priorities of the education system. National and/or sub-national (e.g. state, district) education priorities are often explicit in legislation and regulations as well as policy documentation and statements made by policy makers. The education sector plan is particularly useful in providing an outline of overall education sector goals, cross-cutting issues and planned methods for achieving and monitoring goals—of which an assessment of learning outcomes may be one. The targets outlined in SDG 4 are also an example of education priorities.

**Engage stakeholders.** Involving stakeholders in key assessment decisions helps to improve the usability of assessment results. To involve major stakeholders in important assessment decisions, structures and processes need to be put in place. One way of achieving stakeholder engagement is to appoint a steering committee. The steering committee should comprise both major education stakeholders as well as senior officials and assessment agency staff representing the assessment's overall and technical management. This will help ensure that the overarching assessment design and the required financial, human and physical resources are aligned to successfully and efficiently implement the assessment. At the end of the assessment program the steering committee may meet to discuss the major implications from the

assessment results and to provide guidance around reporting and dissemination strategies. The steering committee should meet regularly to ensure continuity of direction in the assessment program.

**Identify how results will be used.** Understanding how the assessment results will be used helps to define the purposes of the assessment. If results will be used to inform education policy, they have the ability to inform all stages of the policy cycle: results can monitor and evaluate policy; inform the policy agenda by creating awareness about issues in the education system; inform policy formation by identifying characteristics of successful systems; and inform the way a policy is targeted and implemented on the ground. Results can be used to compare individual performance to other individuals or to situate them within groups. Other uses could be to select individuals for advancement to the next level, or determine growth in learning between two levels of schooling.

**Evaluate the feasibility of implementing an assessment based on the policy goals and measurement priorities.** As mentioned, the content, design and scope is directly linked to the policy goals and measurement priorities. Evaluating the feasibility of the implications that a specific policy goal or measurement priority will have on the content, design and scope of the assessment will help to clarify whether the resources available will achieve the objectives of the assessment. If not, resources may need to be adapted or increased. Or, certain goals and priorities may need to be prioritised to better align with available resources. Resources to consider include budget, time, expertise, assessment team and physical resources (see Key Area 2). Design implications include:

- those that affect participation: target population, sub-populations of interest, sample coverage (see Key Area 9); languages tested (see Key Area 7);
- those that affect the content tested: learning domains to be assessed (see Key Areas 4 and 5); contextual data to be collected (see Key Areas 4 and 6);
- those that affect logistics and timing: technical standards (see Key Area 3); assessment cycles, mode of delivery (e.g. computer based, paper based, orally delivered), test design (see Key Area 8); data collection and data processing (see Key Area 10, 11, 12 and 13); personnel needs (see Key Area 10);
- those that affect communications: reporting and dissemination of results (see Key Area 14.

## Key Area 2: Establishing and managing an assessment team responsible for designing and implementing the learning assessment

### Objective

An assessment team with dedicated staff that is appropriately skilled and adequately resourced to respond to the diverse demands of designing, implementing, analysing, and disseminating the outcomes of the learning assessment.

***An assessment team…***

Developing and implementing an assessment program requires adequate staffing, resources, and management. Local cultural, political and financial contexts influence the decisions around establishing an assessment team, however, overarching responsibilities and roles are largely consistent across different countries.

***…with dedicated staff…***

It is essential that staff is dedicated to an assessment program to ensure availability when required. The number of staff will depend on the purpose and scale of the assessment. The staff could be a mixture of full-time, part-time, or contract-based personnel located within an assessment centre or in other agencies. While it is important to clearly define the different roles and tasks within the assessment team, it is equally important that the managers and team members work together across different phases of the assessment program.

***…that is appropriately skilled and adequately resourced…***

Ideally, key management staff have previous experience in educational assessment. However, staff will also be required to develop new skills and expertise. Even very experienced and technically expert assessment teams need to continually develop new skills to keep ahead of new developments, for example, in statistical and measurement theories, or in new technologies and tools. Hence, capacity development is a core component of any assessment team.

Physical infrastructure that is well-fitted to the assessment tasks will be needed. The quality of the infrastructure will depend on the availability of resources and the budget. For this reason, infrastructure needs should be planned at the beginning. When budgeting, experts in the relevant fields should be consulted about specifics—for example, data analysts about statistical software.

***…to respond to the diverse demands of designing, implementing, analysing, and disseminating the outcomes of the learning assessment.***

Large-scale assessments involve an enormous amount of organisation of processes, staff and logistics. There are also often very demanding timeframes involved, with many finish-to-start task dependencies. For example, sampling cannot begin until sample frame information is gathered. Additionally, specific skills in an assessment program are only used for parts of the time. To balance human resources with efficiency and budget requirements, staff could be responsible for multiple aspects of the assessment. Consideration should also be given to outsourcing some aspects of the program or appointing short-term positions from within the agency.

## Achieving the objective

**Establish an assessment team.** The way that the assessment team is organised will depend upon the institution in which it will sit, the funding source for the assessment, and any organisational or bureaucratic constraints present. An assessment team may be located within a ministry, a research agency, a university, or another organisation. It is highly recommended that a core team be established to manage the assessment program on a full-time basis.

Figure 2 shows an example of how an assessment team could be organised. It indicates the core areas of expertise required, and the roles and responsibilities in each area. Project management will involve oversight of assessment development expertise, analytical expertise, operations expertise, and communications expertise. Local factors will greatly influence the organisation and makeup of the assessment staff. Expertise can be extracted from other areas, or external partners where appropriate, for any of the responsibilities outlined. Dependant on the infrastructure of the assessment centre or agency, team members may have multiple responsibilities. During the assessment planning stage, to ensure efficient use of available skills, it is important to consider staffing needs based on assessment tasks and timelines. Since many tasks do not run continuously for the length of an assessment, for example test development (see Key Area 5), consideration should be given to assigning multiple tasks to team members.

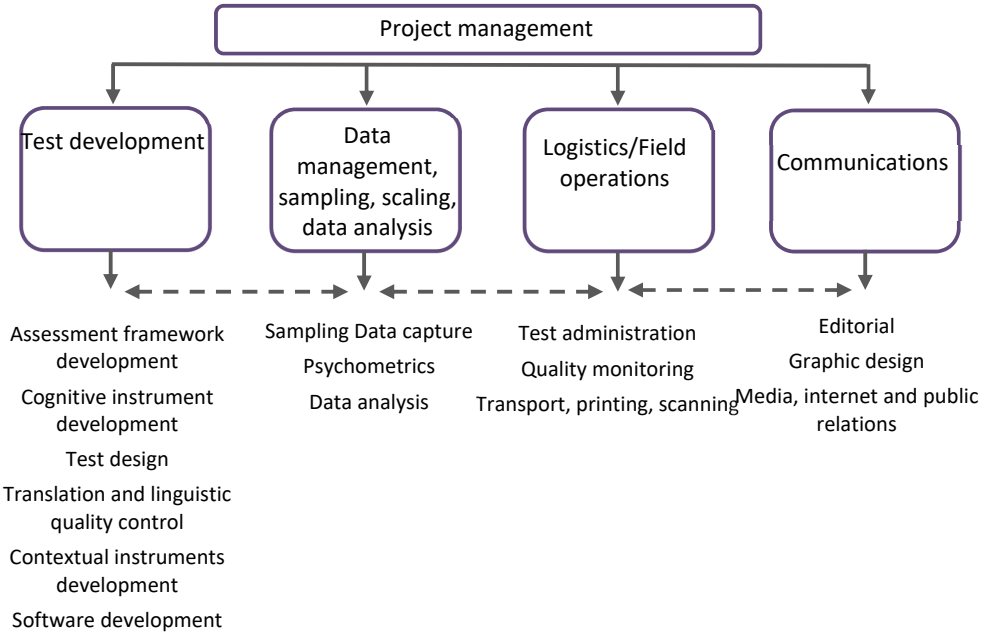


**Figure 2: Example assessment team organization**

**Develop capacity where needed.** Capacity development of assessment staff is a key factor. At the beginning of the assessment program and at regular times throughout, undertake an existing capacity analysis, and a capacity needs analysis. The degree to which there is a mismatch between these two analyses indicates the extent of the capacity development needs. The GP-LA presents principal aspects of good practice in learning assessment and can serve as a basis upon which to analyse the capacity needs of the assessment program.

Once it is clear what capacity needs to be developed within the assessment team, appropriate learning activities should be planned. The 'appropriateness' of the learning activity may depend on its availability, cost-effectiveness and timeliness. Activities could include: study visits; internships; short courses; long courses (e.g. postgraduate diploma); and customised technical training provided by consultants. There are various instances where many staff members may require training in the specifics of the assessment program (e.g. translators). In this case, core staff should conduct the training and develop the training materials.

**Outsource when needed.** At times, it may be more efficient or cost effective to outsource specific tasks to specialist agencies. The assessment team may need support, for example, with item writing, translation, linguistic quality assurance, graphic design, sampling, data processing, psychometric analysis, editing or publishing. Or, the scale of the task may require the assessment team to procure services in printing, distribution, or scanning of assessment materials. Advantages of outsourcing include: access to skills and facilities not available in the assessment centre or agency; increased capacity to process work at peak times; and fixed price and timeframe for delivery. Costs and risks include: time needed to specify criteria, identify contractors, obtain quotes and negotiate price; management of contractor delivery; intellectual property security; and outgoing budget costs. A transparent process should be used when identifying outside contractors.

**Secure physical infrastructure.** Infrastructure to consider includes not only the workspace, but IT infrastructure (servers, networked hard drives, internet connection, cloud access/storage, personal computers, a website about the assessment program), software (security software; standard office communications and support software; database software; desktop publishing package; data management software; statistical analysis package; and scaling software), and other infrastructure including printer(s); a scanner(s) and photocopier(s); telephones, including a dedicated hotline; a storage site for assessment materials, secure from theft, fire and flood; and a secure space for conducting the field trial and main survey data processing and scoring operations.

## Key Area 3: Formulating and articulating technical standards to guide assessment activities

### Objective

A key document that clearly describes standards of technical quality for all aspects of the learning assessment, and indicates how standards can be used as part of quality monitoring and reporting.

***A key document that clearly describes standards of technical quality…***

The term 'standards of technical quality' refers to a set of guidelines that establish the qualities or requirements of a given process or output. In the context of large-scale assessments of learning, technical standards explicitly indicate the expectations on the quality of the assessment endorsed by the governing body of the assessment program such as the steering committee (see Key Area 1).

***…for all aspects of the learning assessment…***

In general, technical standards are established around a set of program components or tasks that relate to the major processes, milestones and deliverables specific to an assessment program. In major assessment programs, the components can be broadly categorised into two groups: 1) the quality of the data; and 2) the management and operations aspects of the program (OECD, 2015).

***…and indicates how standards can be used as part of quality monitoring and reporting.***

The formulation and evaluation of technical standards is a recognised way of assuring the quality of an assessment program. Documenting the adherence to, and deviation from the technical standards will provide transparency on the assessment quality, and thus improve the confidence of stakeholders in the assessment results and support the use of the assessment data in education policy making and planning. The documentation and evaluation against the predefined standards provides evidence of the technical robustness, and enhances the reputation of the assessment program.

Additionally, technical standards provide a basis upon which to develop specific instructions and training activities for various aspects of assessment implementation (Cresswell, 2017b). Where relevant, sections of the GP-LA will outline what type of information should be included in these implementation instructions. For example, Key Area 10 provides details about developing instructions for standardising test administration.

## Achieving the objective

**Establish a technical advisory committee.** This technical advisory committee could be part of the steering committee or a separate entity. In either case, the committee should be made of highly qualified experts who are experienced in the field of large-scale educational assessments. This committee considers the stated policy goals and measurement priorities (Key Area 1) and recommends the levels of precision in various aspects of assessment implementation that are needed to yield data that are useful (Cresswell, 2017b).

**Determine the details of the technical standards.** Depending on the scale of the assessment program, the scope of technical standards and its level of detail can vary. Large-scale international assessment programs usually have a very comprehensive set of technical standards, while small regional assessments may have a set of technical standards that covers a limited number of specific assessment tasks. The following is a set of technical standards that are applicable to large-scale assessments of learning, representing three major components that are common across national, regional and international assessments.

*Sampling.* Technical standards relate to the level of precision and validity of the sample. See Key Area 9 for more details about sampling considerations. Common sampling standards set requirements on the likes of sample size (e.g. the required number of sampled units in the assessment to achieve a predefined level of precision), response rates (e.g. the required percentage for school and student participation) and sample coverage of the target population (e.g. the required percentage of the target population; predefined percentage of allowed exclusions) as protection against bias.

*Data.* Technical standards relate to aspects of the assessment implementation that directly concern the quality of the data. See Key Area 10 for more details about data collection, and Key Area 11 for details about data capture. Common data standards address standardised procedures for the *test administration* (e.g. timing of the test session, seating arrangements, assigning assessment material, etc.), *test security* (e.g. confidentiality of assessment instruments and data), and *data capture* (e.g. standards for indices of inter-scorer agreement, recruitment and training of data entry staff, and data entry audit).

*Psychometric.* Technical standards define quantitatively the technical quality of the test data and the interpretations that can be derived from the test results. There are several psychometric considerations such as reliability, validity, item difficulty, item discrimination, fit statistics, and differential item functioning. See Key Areas 12 and 13 for more details about data analysis considerations and the section titled Key Quality Concepts for Learning Assessment for more detail about reliability and validity.

# Key Area 4: Developing an assessment framework

## Objective

A document that uses a consistent terminology to communicate the purpose and characteristics of the learning assessment to individuals/groups who are working on it and to a broader audience.

***A document that uses consistent terminology to communicate the purpose and characteristics of the learning assessment…***

This document, commonly called an assessment framework, underpins the validity of the assessment by making explicit the aim of the assessment, and what it will cover in terms of content, skills, knowledge, and context. It clarifies how the stakeholders' purposes can be met (see Key Area 1), and provides accountability criteria for the assessment work. The assessment framework defines terms relating to the assessment, which means that when people discuss the assessment, they can communicate its purpose and characteristics clearly.

***…to individuals/groups who are working on it…***

Those contributing to the development of a framework may include academic experts, curriculum authorities, other education agencies, education practitioners, policy makers, and assessment and evaluation experts. It is valuable to bring together people with different perspectives, to draw on a wide range of informed opinion. This will help to ensure that the framework has authority, wide acceptance, and ultimately, an impact on improving learning.

For test and contextual instrument developers, the assessment framework provides a guide for item writing, by explaining what areas the test/ contextual instrument should cover, how the areas are defined, and in the case of tests, what proportion of the test should focus on each area. At a later stage, the framework can be used to evaluate the final test/contextual instrument against the intended aims and coverage.

***…and to a broader audience.***

A framework helps those in the wider community who are interested in the assessment to understand what the assessment is about. It also helps them understand what the assessment results mean.

## Achieving the objective

**Establish expert committees.** Framework development should be led by a group of experts from a range of backgrounds including experts in the academic subject or learning domain (e.g. university professors), experts in large-scale assessment development (e.g. assessment

development manager), and those with subject-area pedagogical knowledge (e.g. teachers) (Mendelovits, 2017). If possible, expert committees should be established for each of the learning domains assessed. As regards framework development for the contextual instruments, a separate committee should be established which contains experts in contextual instrument design, cross-cultural survey design and methodology, and members from the learning domains committees (Lietz, 2017). Test and contextual instrument developers should be included in the expert committee, or attend committee meetings for two primary reasons: so they can provide technical and practical advice about what is feasible; and, so they understand the reasoning behind the framework development and can implement it in item writing (Mendelovits, 2017).

**Develop the framework.** An assessment framework describes the purposes of an assessment and the assessment content, i.e. the learning domains, in detail. As such it provides a common understanding of what the assessment is about and aims to achieve. The assessment framework is also a key guiding document for item development (Key Area 5).

Descriptions of the domain typically include an explanation of the structure of the domain into strands and sub-strands (e.g. reading literacy can be structured into the strands of decoding and comprehension), and a description of the cognitive skills that are intended to be measured in the assessment. Moreover, the assessment framework indicates design specifications such as response formats, structure and test design.

Ideally the framework also includes considerations about the contextual information to be collected as part of the assessment program, for example the underlying policy priorities to be addressed, a description of the main constructs, as well as the theories and models of making associations between contextual information and learning outcomes.

The assessment framework is also a good place to describe the kind of data an assessment will be able to provide, and how data are going to be reported – for example if the data are scaled and if proficiency levels are established.

**Consult stakeholders.** Other stakeholders should also contribute to refining the framework. This could include presenting it to a steering committee in which various stakeholders are represented (see Key Area 1), and/or aiming for a larger audience of policy-makers and learning domain experts. By ensuring that various stakeholders have had the chance to comment on the details of what is assessed, it is more likely that results will be accepted and used in improving outcomes for learners. However, responsibility for finalising the framework should rest with a combined team of experts (expert committee) and test developers.

**Review the framework.** Ideally, work on a framework begins before any assessment items are written so that the framework is available in draft form as a guide to test and contextual instrument developers. However, it is also important that a framework is seen as a work in progress: flexible enough to be modified, in discussion with various audiences, and as results from piloting. Assessment frameworks typically develop over time, with early versions setting down the fundamental components of the assessment that are then validated, revised and further explained as the assessment program matures.

# Key Area 5: Developing high quality cognitive instruments

## Objective

Cognitive instruments containing items with proven reliability, validity and fairness with regard to the population(s) of interest.

***Cognitive instruments…***

The term 'cognitive instruments' (often called a test) in large-scale assessment refers to a set of items used to collect information about what the participant knows, understands and can do in a particular learning domain, or domains. The cognitive instrument is generally composed of instructions and practice questions as well as the test stimuli and related questions/tasks. It is distinguished from contextual instruments which collect information about the personal characteristics, background, attitudes and values of the participant (see Key Area 6). High quality cognitive instruments are the foundation of good assessment—without them, no amount of good practice elsewhere will be sufficient.

***...with proven reliability, validity and fairness…***

Reliability refers to the consistency and accuracy of test and contextual instrument measures and results over replications of the testing procedure (American Educational Research Association et al., 2014). Reliability in a test and contextual instrument is critical as it ensures that the data are precise enough to be used to make decisions on policy, or on classifying individuals (for example, assigning a pass or fail mark).

Validity refers to the extent to which the assessment instruments (usually the tests and contextual data collection instruments) measure what they claim to be measuring for a specified population. Secondly, validity refers to the interpretations made from the analyses and statistics, i.e. that these are correct and appropriate for the proposed use of the data (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards in Educational and Psychological Testing, 2014).

Fairness means that all participants have the same opportunity to demonstrate the skill that is being assessed so that some groups are not advantaged or disadvantaged by factors that are not relevant to the learning domain.

***...with regard to the population(s) of interest.***

The population(s) of interest is defined initially by the policy goals and measurement priorities of the assessment stakeholders (see Key Area 1), and later refined in the sampling plan (see Key Area 9). The cognitive instrument must be designed specifically for this population—considerations include the appropriateness of the content and language level of the questions/tasks, and the range of expected abilities within in the population.

## Achieving the objective

**Establish a team of test developers.** Test developers are those responsible for producing test content and will participate in all aspects of test development described below. Good item writing is a challenging task that requires specialised skills including technical knowledge about writing items for large-scale assessments, knowledge about the test population and the

learning domain being assessed, and the ability to use language in a precise manner, to name a few. For an institution that needs to deliver regular assessments, it is advisable to build a team of dedicated test developers. Alternatively, the task of item development can be outsourced to an organisation that specialises in assessment.

**Establish mechanisms for obtaining input from outside the test development team.** The learning domain expert committee (see Key Area 4) will be essential in providing feedback on the test items at various stages of their development and review. Additionally, asking various other sources, including those closest to the test takers (e.g. teachers), can be helpful as this can reassure stakeholders that items are customised to the test population. Gaining input, in terms of item review feedback, from stakeholders contributes to a fair and inclusive assessment. It is essential, however, that the test development team is involved in the revision of items based on stakeholder feedback and that this team undertakes the final review of all items to ensure technical quality and the required balance of tasks as prescribed in the assessment framework (Mendelovits, 2017).

Write test items. The assessment framework provides critical information to guide item writing, including the range of content, strands and sub-strands, and skills to be assessed, the different response formats to be used, and the length of the assessment and duration of the test.

Item writing starts with identifying or creating appropriate stimulus material, which is the prompt or context on which an item is based. The next step is to formulate the tasks that relate to the stimulus. An important consideration for this step is about the appropriate response format (e.g. multiple choice or open-ended format). The response format should take into account the nature of the learning domain and how the response formats can validly capture information about the participants' knowledge, skills and understanding in that learning domain. It is essential that when developing the tasks, the responses are developed and formulated at the same time. While this seems to be a natural step for developing multiple choice items, it is also highly relevant for open-ended questions, in order to anticipate the kind of responses participants' may give (see 'Write scoring guide' below). Another important consideration is about the item difficulty. If the intention is to report on the spread of achievement across a population, the items need to cover a wide range of difficulty around the expected ability levels of the participants. The specific item characteristics are essential to be described as part of the metadata for each item (see Key Area 7) – mainly the specifications outlined in the assessment framework, for example the strands and sub-strands, the item descriptor (or 'question intent') that describes what the item intends to measure, the item difficulty, etc.

More items than are needed in the main survey should be developed as a substantial number may be dropped after the field trial. Where a test is to be translated into more than one language, the test materials should not pose any linguistic problems when translated (see Key Area 7). Where a test is to be delivered on a computer or another device, cooperation with software developers will ensure that items meet technical specifications.

**Develop scoring guides.** Scoring (or 'marking', or 'coding') describes the process of classifying responses and allocating codes to represent the various categories of response. These codes are

then allocated scores corresponding to their quality in preparation for data analysis. A scoring guide describes how to score the participants' responses to an item or a collection of items – irrespective of the item format. Thus, a scoring guide should be developed for each item, during the item development process. This includes giving a precise, succinct description of each response category and providing example responses for each category that satisfy each description. In addition, the scoring guide should contain a general description of the skill/process/knowledge the item intends to assess; list the item characteristics as given in the assessment framework; and give a justification of the correct response and distractors (for multiple choice items). The scores for each item should be determined by the test development team in consultation with the psychometric analysis team, and the codes and scores should also be recorded in the item codebook (see Key Area 11). Automated, computerised scoring processes need to be designed and thoroughly checked for accuracy and reliability.

**Develop instructions.** Cognitive instruments should contain instructions for the participants on how to complete the test, along with some practice items. Participants must clearly understand how to respond to the test items, and sufficient time should be allocated before the test to ensure this is the case.

**Review, pilot and field trial test items.** The initial development of an item is just the first step in the item writing process. Items should be reviewed at several stages of the program. During the item development process *panelling* should occur in which test developers and subject matter experts iteratively review and revise items for improvement. It is possible that an item or set of items may go through several reviews and revisions of this kind. *Piloting, or cognitive pre-testing* should usually occur after early rounds of panelling and administers the items on a small group of participants similar to those targeted in the test. Once the items have been piloted and subsequent rounds of review and revision have occurred, the items are tested on a larger scale to assess their validity statistically. The goal is for the final set of items to be statistically sound and to reflect the balance of content and difficulty levels etc., as outlined in the assessment framework. The *field trial* is crucial in establishing the technical robustness of the test. The necessary size of the field trial sample should be determined in consultation with sampling and psychometric experts to make sure there are sufficient data per item to undertake meaningful item analysis (see Key Area 9).

# Key Area 6: Developing high quality contextual instruments

## Objective

| |
|---|
| Contextual instruments containing items with proven reliability, validity and fairness with regard to the population(s) of interest. |

***Contextual instruments…***

The term 'contextual instruments' refers to a set of items used to collect information about the personal characteristics, background, attitudes and values of participants in their contexts (e.g. home, classroom, school). Typically, large-scale assessments use questionnaires to collect a wide range of information from participants. Other forms of contextual data collection include interviews and observations, as for example in household-based assessments. While the

cognitive instruments collect data about the participant's performance in a certain learning domain (see Key Area 5), the contextual instruments collect data on important indicators for reporting, for example gender, age, or socio-economic status, as well as indicators of learning outcomes such as attitudes, values, and behaviours. The contextual indicators are also important as they can be associated with or account for differences in performance.

A good contextual instrument collects accurate information that is relevant to policy makers in areas they can affect or influence, and that research indicates is related to performance (Anderson & Morgan, 2008).

***...with proven reliability, validity and fairness…***

In relation to contextual instruments reliability means that contextual items are able to elicit accurate information from the participant. For example, an item may be unreliable if the respondent is likely to give a socially desirable answer rather than a truthful one. Validity means that the contextual instrument collects the data needed to address the purpose and goals of the assessment, and that questions have a similar meaning for all participants (Lietz, 2017). Fairness means that the contextual instruments are appropriate for all participants.

***...with regard to the population(s) of interest.***

The population(s) of interest is (are) defined initially through the policy goals and measurement priorities of the assessment program (see Key Area 1) and is later refined in the sampling plan (see Key Area 9). While the policy goal may be to assess performance of students in Grade 6 for example, the contextual instrument may need to gather information not only from students, but also from their parents, school principals, or teachers. Ensuring each contextual instrument is appropriately designed in terms of content, participants' reading ability, and cultural considerations is essential for collecting high-quality contextual data.

## Achieving the objective

**Establish a team of** contextual instrument **developers.** Contextual instrument developers are those responsible for producing contextual content and will participate in all aspects of contextual instrument development. Good contextual item writing requires appropriate technical skills regarding question and response scale wording, design and order, precise language use, awareness of translatability issues, a knowledge of the target population, and familiarity with previous research on which contextual factors are relevant to learning in the domain of interest.

**Establish mechanisms for obtaining external input and feedback.** The expert committee for contextual instrument development (see Key Area 4) will be essential in providing input and feedback on the contextual instruments at various stages of their development. Additional input and feedback, for example, for verifying the policy relevance of the contextual instruments' content might be gained from various key stakeholders. This could include the assessments steering committee (see Key Area 1), or other important stakeholders in the assessment.

Input tends to be sought at the following stages of item preparation (Lietz, 2017): at the start of the contextual instrument development to provide sample questions; once the field-trial

instrument has been developed to comment on implementation of actual questions in the survey population; after the analyses of the field-trial data and prior to the finalisation for the main survey to assist with item selection.

**Write contextual items.** The assessment framework is an important tool to guide contextual item writing, including the underlying policy priorities to be addressed, and a description of the main constructs (see Key Area 4).

To get started, it is helpful to consult contextual instruments from other assessments that have a similar purpose. When formulating contextual items, it is essential to keep the language simple, brief and concise. Deciding on the response format and scale is an important step in contextual instrument development, and the formats and scales chosen should be adequate for the kind of information sought. Each item should have an instruction on how to respond. Instructions should be clear and concise, and be the same for recurring item formats. In case of sensitive content, or content that may elicit a socially desirable response (e.g. questions related to socio-economic status, attitudes, values, and behaviours), it is recommended to consider whether the participant will be able or willing to give a honest answer to the question.

Where a contextual instrument is to be adapted to different contexts (e.g. for different year levels), or translated, it is recommended to provide adequate instructions during development (see Key Area 7). This helps to ensure the contextual items maintain equivalent meanings across contexts and languages. Where a questionnaire is to be delivered on a computer or another device, cooperation with software developers will ensure that items meet technical specifications.

**Review, pilot and field trial contextual items.** Like for cognitive items, contextual items should be reviewed at similar stages in the development process. A first step is to pilot contextual items among the developers, by completing the questions as if they were a respondent. Any instance in which there is uncertainty or language which they think can be improved should be noted and addressed in revisions of the questions. Once the team is satisfied with the instrument, it should be distributed to the relevant stakeholders in the assessment program for their feedback. Piloting the contextual items to a small group of participants is crucial. In particular, respondents can be asked about their impressions of the comprehensiveness and applicability of the questionnaire/interview/observation content as well as any items they found confusing or unclear. This qualitative information can also inform further revision of the items prior to commencing field trials.

## Key Area 7: Linguistic quality control for translation of cognitive and contextual instruments

### Objective

| |
|---|
| Cognitive and contextual instruments that are appropriate, linguistically equivalent, and psychometrically equivalent across multiple languages. |

***Cognitive and contextual instruments that are appropriate…***

'Appropriate' means that assessment items are sensitive to the cultural differences of all participants taking the assessment and to the structural differences of the context. For example, if the assessment is to be conducted in rural and urban areas, certain topics like metropolitan rail networks may not be appropriate for rural participants, and vice versa. Structural differences refer, for example, to different study programs that are available, different forms of grouping within schools (e.g. based on achievement), school shifts, etc. It is therefore recommended to develop and include adaptation guidelines to ensure cognitive and contextual instruments are appropriate for different cultural and structural contexts, while at the same time ensuring the content is equivalent.

***…linguistically equivalent…***

The term 'linguistically equivalent' means that two or more language versions of the same instrument have the same meaning. In some cases, maintaining the same meaning may require some adaptation of the item text. For example, when indicating that a response lies between 5 and 10, the word 'between' may include 5 and 10 in one language, while in another it might exclude 5 and 10 (Dept, Ferrari, & Halleux, 2017). The item therefore needs to be adapted by adding further specification to ensure participants of both languages are including 5 and 10 when considering their response.

***…and psychometrically equivalent…***

The term 'psychometrically equivalent' means that the two language versions of the same instrument are equally difficult for participants of the same ability level across languages. For example, if the wording of an item in the source language is clearer than the wording for the same item in a target language, then participants completing the item in the source language would find it easier to answer than those completing it in the target language. The translation should therefore be adapted to make the wording of the target language clearer.

***…across multiple languages.***

If the assessment is conducted across multiple languages, it is likely that one of its purposes is to make comparisons between test language groups. For the comparison of results to be valid, assessment agencies need to be certain that results showing differences between languages are due to factors inherent in the participants and not the assessment instrument itself. Ensuring that items are culturally appropriate and linguistically and psychometrically equivalent across various language versions of the assessment instruments helps to strengthen the validity of these comparisons.

## Achieving the objective

**Develop a translatable source version.** Translations are based on one (or in some cases, two) initial language version (or a 'source version') of the assessment. To help keep linguistic quality control processes efficient, translation issues need to be considered throughout the source item development process. To facilitate this, test developers should be trained to be aware of common translation issues. For a subset of items, an evaluation by a linguistic expert of how amenable the items are to translation (often called a translatability assessment) can further identify issues specific to the testing languages. This information can feed back into training

and inform translation guidelines. Ensuring source versions are translatable before they are finalised will avoid the need to make changes later in the process—changes that increase the chance of introducing error.

**Provide supporting documentation.** Documentation supports a number of translation tasks: general translation guidelines provide an overview of the assessment, the importance of linguistic and psychometric equivalence, and general assessment-related translation issues; item-specific guidelines include clarifications about the meaning of words or phrases and acceptable or unacceptable adaptations; in addition, a glossary helps to ensure consistency of terms across items and translators.

**Recruit and train translators.** Translation is a complex and nuanced task that requires individual judgement based on experience and skill. As such, translators should be formally qualified with professional experience. Ideally, translators are familiar with the subject being assessed and the principles of good item writing. Training consists of familiarising translators with the assessment, the workflow, and conducting translation exercises.

**Implement a multi-step process.** Translation is a subjective task. Therefore, to increase the accuracy and equivalence of the assessment, translations should be reviewed by a number of experts. A forward translation workflow is considered the best way to manage this. In a forward translation workflow, judgements about the quality of the translations are made in the target language version. Back translation is another workflow commonly used in assessment. However, since items are reviewed in the source language rather than the target language, some errors can be missed by the reviewer (Hambleton, Merenda, & Spielberger, 2004). For example, errors in fluency or register may not be exposed in a back translated version of the item, while they will be evident in the target version.

Judgements about the quality of the translations require various experts. Issues related to linguistic equivalence of the source version and the translated version should be carefully reviewed by other translators or linguistic experts. Issues related to psychometric equivalence can be reviewed by item developers or other learning domain experts. All changes in the translated/adapted instruments should be recorded centrally for future reference.

**Manage the workflow.** The forward translation workflow requires careful management as various experts are involved at various stages, and each expert's work depends on the previous experts completing their task. Sometimes the source version needs to be updated, requiring translations to be revisited and updates recorded. Clear management processes, including timelines and documentation of decisions, need to be in place to ensure sufficient time for thorough quality control and to avoid the introduction of errors.

**Analyse field trial statistics.** Even when a robust forward translation workflow is completed, linguistic and psychometric equivalence cannot be confirmed until item statistics are examined. A field trial (or 'pre-test') of the items provides the opportunity to gather data and examine the items statistically before finalising them for use in the assessment. Statistical tests of equivalence compare results across language groups and identify potential problems. Problems should be investigated to determine whether translation is the cause. If translation is the cause,

a decision should be made as to whether the item is included in the main survey, modified and included, or removed from the assessment completely.

## Key Area 8: Designing the cognitive and contextual instruments

### Objective

A design that ensures efficiency in sample sizes, balanced assessment content, appropriate assessment length, and stable measures over time.

***A design…***

The 'design' of cognitive and contextual instruments refers to the allocation of items within a test, questionnaire, or interview, and across multiple forms of these instruments. The term 'test design' refers to the allocation of items within and across test booklets in order to achieve full coverage of the learning domain specifications defined in the assessment framework. Similarly, 'questionnaire or interview design' refers to the allocation of contextual items within and across questionnaire/interview forms. Depending on the purpose of the assessment and the number and kind of contextual questions asked, such items could be administered as part of the cognitive tests (e.g. as a separate section of the test booklet).

***...that ensures efficiency in sample sizes…***

If the major purpose of the assessment is to compare performance at the individual level, then the assessment will likely take a census of the population and all participants will take essentially the same test. However, if the purpose of the assessment is to describe performance across the education system as a whole (or a clearly defined part of it), then testing a sample of the population may suffice and, in order to test a broad range of curriculum or learning domain content, tests and questionnaires/interviews may be split over several versions. The amount of content and the way in which it is distributed across test and questionnaire/interview forms will affect the size of the sample of participants needed to provide precise data (see Key Area 9).

***...balanced assessment content…***

Whether the test and questionnaire/interview uses multiple forms or just one, the distribution of the content in each form should adhere to the blueprint as described in the assessment framework (see Key Area 4). When using multiple forms, the test and questionnaire/interview design should allow for comparable measures of achievement or contextual information to be established regardless of which form a participant completed.

***...appropriate assessment length…***

It is important to plan the amount of cognitive material allocated to each test form and to design forms that can be reasonably completed in the time allotted. If time allocated to complete the test is too short then the individual will be working under time pressure which may introduce a greater risk of guessing. If a test is too long (there are too many items) then fatigue may have an effect on achievement. Appropriate length is also an important consideration for contextual instruments in order to keep respondents engaged.

***...and stable measures over time.***

If the purpose of the assessment is to measure growth over time, a linking design should be used. A linking design uses a common set of items across years or grade levels. This enables the data for each learning domain collected from the various assessments to be statistically linked and combined onto a single scale.

## Achieving the objective

**Consult psychometricians and item developers.** The development of the design of the cognitive and contextual instruments should be done in consultation with psychometricians and item developers. Test and questionnaire/interview developers will ensure that the design matches the blueprint specified in the assessment framework. Psychometric experts will be able to advise on the consequences that any design will have on the ability of the test and questionnaire/interview data to inform the policy goals and measurement priorities of the assessment.

**Allocate items within and across test forms.** If only one test form will be used for all participants, the major concern is that the form is structured so that it begins with some easy items to encourage lower-ability participants to attempt the test. It is then often desirable to mix the difficulty of the following items so that participants do not abandon the test when they encounter a run of difficult items. It is also important to have some harder items earlier in the test so that participants who work more slowly have an opportunity to attempt some harder items. As lower-ability participants are less likely to finish the test, it is good to end with some harder items to challenge those participants who can reach them.

If several test forms will be used, the test design becomes more complex. In order to establish a common scale for each learning domain in a test, each test form should be linked to another using a *linking design,* i.e. a common set of items. The link items should display good statistical quality and, as a set, should adhere to the assessment framework blueprint as closely as possible. It is also important that the link items function in the same way across forms and across different groups of participants (e.g. the overall difficulty of each form should be comparable). Thus, it is essential to determine the psychometric equivalence of the test forms as part of the field trial and main survey.

Linking designs are also used to establish growth over time, and across grades/year levels. Historical linking achieves this by using items from previous test administrations in the current test for the same population, e.g. grade 10 students. Vertical linking achieves this by using items from a lower grade level at a higher grade level. To ensure comparability over time, the set of items selected for linking should remain the same.

**Allocate items within and across questionnaire/interview forms.** Questionnaires or interviews are usually administered after the cognitive test, which means that by the time the participant starts the questionnaire/interview, she or he could be tired. As such, the structure of the questionnaire/interview form should place the most important questions at the beginning to ensure the highest chance of them being answered. Important questions frequently include demographic questions regarding gender, age, language spoken at home and socio-economic background factors as these are important indicators for reporting and in association with

achievement (e.g. difference between girls and boys in reading or mathematics achievement, strength of the relationship between socio-economic status and achievement).

In terms of ordering content related questions, general questions should precede specific questions as the latter tend to take 'parts' out of the former. Thus, in a study of student well-being, for example, students should be asked about how they feel in general first before they are asked about how they feel at school.

Rotation designs are commonly used for questionnaire forms in the field trial. By using a rotation design in the field trial, it is possible to review a larger number of items without increasing completion time for participants. Rotation designs in questionnaires include a set of items that are common to all forms and that represent core contextual factors like gender, language spoken at home, etc. Rotation designs then have a number of sets of items rotated across forms. Rotation designs are less commonly used in main surveys because they may impact the strength of the data at the school or village level (Lietz, 2017).

**Allocate forms to participants.** If a number of rotated forms will be used, then a form allocation scheme should be established. Forms need to be distributed to participants at random but at the same time, they must be distributed evenly. That is to say that each form should be allocated to a similar number of participants.

**Layout and proofread the forms.** Tests and questionnaires should be visually engaging *without* being visually distracting. The layout should be clear and consistent, suitable for translation (e.g. considering different scripts), and possibly accommodate for visually impaired participants. Sufficient space needs to be provided for participants to write answers to open-ended responses (Anderson & Morgan, 2008).

It is essential to provide room for recording the unique Identification (ID) which should be allocated to each participant in the assessment. The ID is essential in matching participant-level data from the test and/or questionnaire/interview, to the school-/teacher-/household-level data. Another administrative layout feature is to consider adding an item identifier (or item code) next to an item. This allows the item to be identified quickly and easily during checking, scoring, data cleaning, and analysis. The item identifier should be added in an unobtrusive way (e.g. in greyscale).

Proofreading is an extremely important process that should be performed at various stages during instrument preparation (e.g. before and after forms are sent for printing, or computer-based forms have been created). In computer-based tests, content, design and functionality testing need to be performed as part of the layout process.

# Key Area 9: Sampling

## Objective

A sample that, through the use of scientific sampling methods, helps to guarantee appropriate levels of statistical precision and validity in the interpretation of assessment results.

***…A sample that, through the use of scientific sampling methods…***

Scientific sampling methods ensure that the sample is reflective of the population and inferences about the population can be made from observations of the sample. This means that statements about the population can be based on the findings of the study conducted using the sample (Ross, 2005).

Testing a sample of a population is an effective and efficient way to gather information to describe performance across the education system (or a clearly defined part of it). If the purpose of an assessment is to report on the performance at the individual level, then it is likely that a census of the population will complete the assessment. .

***...helps to guarantee appropriate levels of statistical precision…***

The statistical precision derived from any sample depends on 'sampling error' (Bordens & Abbott, 1996). Sampling error occurs when the characteristics of the survey sample are somehow different from the characteristics of the population of interest. If inadequate methods are used to select the sample, and there is a failure to minimise sampling error, the advantages of studying a sample to make generalisations and inferences about the population of interest are diminished (Bordens & Abbott, 1996; Floyd & Fowler, 2009).

There are two main types of error present when selecting a sample: *error due to chance*—because samples are a subset of the population of interest, there is always a chance that the true values resulting from a sample are not the same as the true values of the population (Floyd & Fowler, 2009); and *error due to bias*—some systematic inclusion or exclusion process as a result of the selection method or due to respondents' decision to participate or not participate in the study (also known as response bias) (Bordens & Abbott, 1996; Floyd & Fowler, 2009). Error due to chance can be minimised by increasing the sample size. Error due to bias is reduced by using a probabilistic sampling method in which each population member has a chance of being selected and the probability of selection is known at the time of sampling (Murphy & Schulz, 2006; Ross, 2005).

***…and validity in the interpretation of assessment results.***

Validity is generally known as the extent to which the cognitive and contextual instruments measure what they claim to be measuring for a specified population. Validity in the interpretation of results means that the interpretations made from the resulting statistics are correct and appropriate for the proposed use of the data (American Educational Research Association et al., 2014). The sampling process informs the interpretations in terms of the extent to which results are generalisable to the population. For example, if a sub-group of the population is excluded from testing (e.g. the private sector), it must be clear that results do not reflect the performance of this sub-group.

## Achieving the objective

**Develop a sampling plan and select a sample methodology.** A comprehensive sampling plan and design should be prepared at the beginning of the assessment program, including a comprehensive list of sampling activities and timelines to which they need to be completed.

The sampling plan will also outline the sampling methods to be used and the sample size needed to ensure reliable data. These aspects are described in detail below.

*Define the target population.* The target population is the population of interest whose performance will be described in the assessment results. The definition is made up of two elements: the *desired* target population is the population about which inferences from the survey outcomes will be made, and the *defined* target population is the desired target population minus certain elements that are excluded due to practical difficulties (Murphy & Schulz, 2006). For example, the desired target population may be all students in Grade 6, but the defined target population may exclude Grade 6 students in very remote locations due to the practical difficulty and costs involved in reaching those students. *Exclusions* from the defined target population need to be clearly documented, and the extent of non-coverage (of the population) needs to be estimated using available data sources—e.g. census and enrolment data (Murphy & Schulz, 2006). It is preferable that the defined target population be as close as possible to the desired target population (Murphy & Schulz, 2006).

*Define sub-populations of interest for reporting.* Sub-populations of interest should be defined during the preparation stages of the survey (see Key Area 1). Sub-populations could be based on, for example, gender, social category, urban/rural locations, or districts. During the sampling stage, these sub-populations should be factored into the design of the sample. This is important, as the sample size for each sub-population needs to be large enough to enable statistical comparisons across sub-populations.

*Choose a scientifically sound probability sampling methodology.* Sound probability methods are those in which each unit of the target population (e.g. school, student) has a known, non-zero probability of selection (Thompson, 1992). There are two main types of probability sampling methods: *simple random sampling* and *complex random sampling.* These methods randomly select units and therefore have no systematic bias.

In large-scale educational assessments, however, simple random samples are impractical. For example, a random sample could select 400 students from 400 different schools across the country, making it impractical to go to 400 locations to administer a survey just to a single student. Therefore, a complex random sample using *clustering* is usually used. However, the use of clustering decreases the efficiency of the sample. This means that more participants are required than would be the case under a simple random sampling scenario to achieve the same level of precision.

The choice of sampling method should be made in consultation with sampling experts and will be based on the cost and logistics of data collection, the need to have a sufficient level of precision, and the analytical goals of the survey (Murphy & Schulz, 2006; Ross, 2005).

*Determine an appropriate sample size.* As previously indicated, the sample size will be determined by the sampling method used and the desire to compare sub-populations of interest. It will also depend on practical considerations like costs, timelines, logistics, and the availability of accurate data for constructing the sampling frame.

**Construct or obtain a comprehensive sample frame of the target population.** A sampling frame is a list of all the sampling units for a sample survey. A well-constructed sampling frame is one that provides complete coverage of the defined target population. For example, for an educational assessment survey of Grade 6 students in a particular region, the sample frame will contain a list of all the schools in the region containing Grade 6 students, along with an estimate of the number of Grade 6 students in the school. The quality of the sampling frame has a direct effect on the survey results as it is used to select the sample for the study; thus, sample frame construction should be treated with extreme care (Murphy & Schulz, 2006; Ross, 2005).

Dividing, or *'stratifying',* the sample frame by key sub-populations of interest improves the efficiency of the sample design and ensures that all parts of the population are included and represented in the sample. Stratification variables should be considered during the construction of the sample frame to ensure the required data is collected (Murphy & Schulz, 2006; OECD, 2005b).

**Conduct the sampling.** There are usually two major data collection stages in large-scale assessments, the *field trial* and the *main survey.* Main survey sampling should be conducted first and it should strictly adhere to the chosen sampling methodology. At this stage, substitute units may also be identified in order to minimise the potential for bias if some units do not participate in the assessment.

If both the field trial and main survey sampling are conducted in the same year or cycle, field trial sampling should use sampling units not chosen for the main survey. It does not need to be as rigorous as for the main survey. This is due to the different goals of the field trial which are to test the survey instruments and to test operational procedures. The field trial could be limited to fewer regions to contain costs and to minimise the burden on smaller jurisdictions. It should, however, be approximately representative of the population of interest. The size of the field trial sample is based on the amount of response data required to be able to adequately test the psychometric properties of the survey items, and should be determined in consultation with sampling experts.

**Maintain optimal response rates.** Response rate is calculated as the number of units that participated divided by the number of units that were sampled expressed as a percentage (Bordens & Abbott, 1996). The response rate can be used to gauge the potential for non-response bias, therefore the higher the response rate of a survey, the lower the risk of non-response bias. Every effort should be made to ensure a high number of sampled units participate in the survey. Efforts such as promoting the importance of participation can raise the response rates. Using substitute units can also raise the response rates, however this does not completely ameliorate bias as the actual sampled unit isn't responding and it is only hoped that the substitute would respond in a similar way. Conducting follow-up sessions when response rates of initial test administration sessions are low will assist in raising the participant-level response rates.

**Apply the proper weighting methodology to improve the accuracy of estimates and to obtain correct standard errors.**

*Sampling weights* are used to correct for imperfections associated with the sample that might lead to bias and other differences between the sample and the population of interest. The purposes of sample weighting are (Foy, 2000): to account for the sample design—weighting compensates for any stratification or disproportional probabilities of selection of subgroups; to adjust for non-response and non-coverage (due to errors in the sampling frame) of the population; to adjust the weighted sample distribution for key variables of interest (e.g. age, gender); and to make it conform to a known population distribution. Once the imperfections in the sample are compensated for, weights can then be used in the estimation of population characteristics of interest and also in the estimation of the sampling errors of the survey estimates generated (Foy, 2000).

*Replicate weights* are used to compute standard errors in analysis of complex survey data (OECD, 2005a). In doing this, they place the burden of understanding the complex sampling design onto the people preparing the final database, rather than the data analyst (Rust, Krawchuk, & Monseur, 2017). This is useful because assessment data may be analysed not only by the assessment team, but also by other stakeholders such as policy makers, or researchers.

# Key Area 10: Standardised field operations

## Objective

Field operations that are standardised, documented and monitored to ensure that the data are collected under the same conditions, independent from the administration context, in an efficient and secure manner.

***Field operations…***

The term 'field operations' refers to the activities of administering the assessment, including planning, implementing, documenting and monitoring the data collection 'in the field'. There are usually two major data collection stages involved in large-scale assessments—the field trial and the main survey. The primary aim of the field trial is to evaluate the assessment instruments (see Key Areas 5, 6 and 7) and the operational procedures. As a result of the field trial, any identified issues relating to the assessment tools or to the operational procedures can then be improved and implemented in the main survey.

***...that are standardised, documented and monitored to ensure that the data are collected under the same conditions, independent from the administration context…***

The standardisation of assessment administration procedures is essential to ensure that each participant is assessed under the same conditions. This helps to guarantee that differences in performance are inherent in the participant, and not a result of testing conditions. Standardised conditions include standardised training for all test administrators, common procedures for contacting participants (e.g. through schools), test and questionnaires with similar print quality (in the case of paper-based assessment), the same timing schedule for all participants to

complete the assessment, strategies for ensuring the security of test materials, etc., and are described in more detail below.

Field operations procedures should be described and documented in manuals and other materials as required. Fully documenting field operations is important, not only for those implementing the assessment, but also to reassure and inform the general public of the rigorous procedures undertaken during data collection.

To confirm the integrity and high standards of the field operations process, quality assurance procedures, such as on-site monitoring, should be employed. Quality monitoring of the test administration involves the physical presence of a quality monitor at the testing site where they observe and report whether the assessment is taking place using standardised procedures under similar conditions (ACER, 2005).

***…in an efficient and secure manner.***

Efficiency in field operations procedures ensures costs are minimised – both in terms of budget as well as time and effort for participants and others involved in the assessment administration (e.g. school staff, education authorities, test administrators, etc.)

Security concerns the confidentiality of the assessment material – especially the cognitive instruments, at all times during preparation, administration, and submission of the (completed) assessment material, to allow for linking over time, and to prevent cheating. Security is also concerned with all aspects of the assessment administration that may lead to deviations from the standardised procedures, and may have an impact on the confidentiality of the assessment material or the data collected. For example, any interruptions of a test session, such as presence of uninvited people (e.g. press, or uninvited school staff), or a fire alarm. A deliberate anticipation of such possible events, together with a thorough documentation of any deviations from the standardised procedures are essential quality assurance mechanism to ensure confidentiality of test material, participants and the quality of the data collected.

## Achieving the objective

**Develop an operation plan.** The operation plan should contain a comprehensive list of field operations activities and the timelines to which they need to be completed. The operation plan helps to ensure that the data are collected and are ready for analysis in a timely manner. Another benefit is that the important activities have been listed in the operation plan, thus minimising the risk of accidently omitting activities if adjustments to timelines need to be made. An operation plan may encompass all field operations activities associated with the assessment program, or alternatively, there can be several operation plans that are specific to key personnel (e.g. program managers, test administrators, and quality assurance monitors) or key stages (e.g. assessment material preparation and production, assessment administration/data collection, data processing and data management).

**Prepare field operations manuals and documentation.** Field operations manuals (otherwise known as field guidelines or administration manuals) and important supporting documents are essential tools to ensure standardised field operations. Field operations manuals assist field personnel with the preparation of resources required for the field operations tasks as well as the

completion of the tasks themselves. These can include, for example a test administrator's manual and a quality monitor's manual. Other documentation to be used in the field, for example attendance forms and quality control checklists, are used by field personnel to support and verify the quality and accuracy of the field operations.

It is essential that all manuals and other documents are implemented in the field trial to make sure the procedures outlined in the manuals work, and that the manuals and documents are useful for the key administration personnel. Based on the experience in the field trial, the manuals and documents can then be revised and improved for the main survey, to ensure high standardisation and quality of the assessment administration.

**Contact sampled institutions and individuals.** Contacting sampled institutions and individuals will depend on the type of assessment (e.g. school-based or household-based), and any local conventions and regulations. The cooperation of institutions and individuals is essential to achieving optimal response rates (see Key Area 9). Sampled institutions (e.g. schools, or village councils for household-based assessments) and individuals should fully understand the purpose of the assessment, the importance of achieving a representative sample, what their role is in the assessment, the security measures taken to maintain confidentiality of responses, and any regulations permitting the assessment to take place. If an institution chooses not to participate, replacement institutions should be contacted. For this reason, institutions should be contacted as early as possible before the assessment date.

**Recruit and train key personnel.** Key personnel needed for administering an assessment usually include *test administrators* who are responsible for delivering the assessment to the participants; *quality monitors* who observe the assessment administration and verify that test administrators follow the procedures as outlined in the manuals; and a *contact at the institution* (or village in household-based assessments) who will help manage any requirements at the institution level.

Test administrators may be trained staff from within the school or community, or independent personnel administering the test in various locations. If a teacher administers the test in a school, it is preferable that she or he is not the teacher of the students taking the test (Cresswell, 2017b).

The number of key personnel required depends on the number of sampled institutions or individuals, and the number of test sessions that need to be administered or monitored.

It is important that all key administration personnel are informed about the purpose of the assessment and their roles and responsibilities in its administration. Test administrators and quality monitors need to be trained in administering the test under standardised conditions, and understand why standardised test administration procedures are vital. As such, training should place particular emphasis on uniform testing conditions (ACER, 2013). Training should be conducted well before the data collection, allowing enough time for key personnel to complete any preparatory tasks required for the data collection.

**Check assessment materials.** It is important that quality control procedures are implemented to ensure assessment and administration materials are of high quality. For *paper-based*

*assessments* this includes stages of proof-reading throughout the development of the material, as well as a final check before mass production/printing (e.g. a check of print quality, correct pagination, participant identification). For *computer-based assessments* software and device testing (at the assessment agency and at schools – if school devices are used for test delivery), as well as final checks of the software before mass production (e.g. on USB sticks) or delivery, are essential quality assurance procedures.

**Administer the assessment according to standardised procedures.** It is vital that administration procedures are carefully followed by all key personnel involved in the assessment administration. Administering the assessment refers to the activities carried out at a school or in a household, including the handling of materials, administration of the tests and contextual data collection, organisation of follow-up sessions, quality assurance monitoring, and the collection and storing of the completed materials.

**Verify and register the return of assessment materials.** All assessment materials should be verified for completeness and registered on return. Any missing materials should be followed up, especially if they contain data, such as tests and questionnaires or attendance forms. It is important that all assessment materials are returned before data capture so as to ensure all data collected is available to be recorded in the database, thus securing response rates and adequate coverage of the sample.

# Key Area 11: Managing data

## Objective

A final database that respects respondent anonymity, is free from discrepancies and errors, and is appropriately structured and documented for analysis and dissemination.

***A final database that respects respondent anonymity…***

Participant privacy and anonymity is an essential ethical aspect of large-scale assessments (see *Key Quality Concepts*). Even if the aim of the assessment is to report individual results to the school, teacher, and/or parents, data that are released publicly should be anonymous. Data could be anonymous from data collection, where participant names are never collected, or data could be de-identified before public release.

***...is free from discrepancies and errors…***

Discrepancies in data occur when there is conflicting information for the same fact. For example, the age that a student recorded in the questionnaire, may be different to the age recorded by the school. Errors in data usually occur during data capture[1], where a response has been captured incorrectly, or not at all.

It is essential that quality control measures for data management are embedded in all stages of data management, especially during data capture, data cleaning, verification and validation.

***...and is appropriately structured and documented for analysis and dissemination.***

[1] 'Data capture' refers to manual data entry as well as electronic data processing such as optical character recognition or optical mark recognition.

Learning assessment databases need to be designed for the user. Depending on the program's mandate, particular stakeholder groups such as government officials, researchers, and the general public may need access to the data and therefore specific databases may need to be constructed to fit each user's needs. Databases should be accompanied by documentation that describes what each variable in the database means and, if relevant, how raw data was transformed. Documentation helps, not only the external user to understand and use the database, but it also helps data management personnel to create the database and enter, clean, verify and validate the data.

## Achieving the objective

**Develop a data management plan.** The scope and level of detail of the data management plan will depend on the complexity of the assessment. The plan should be prepared at the beginning of the assessment program, and it should include a comprehensive list of data management activities and the timelines to which they need to be completed.

**Design the codebook.** The data codebook is an important data management tool. It contains the information about all the variables in the database including their definitions, data type, values, validity parameters (e.g. number of characters/decimals allowed for valid values of each variable in the database), and codes for missing values. The codebook format depends on the data management software that will be used.

The design of the codebook specifies how the raw data from the test and contextual instruments are structured into the data file. It should take into account the planned analyses so that the raw data are converted appropriately.

**Prepare data management and data capture software and hardware.** Large-scale data collection is a complex process that can involve a large team and numerous cycles of data collection over an extended period of time. The use of data management software can substantially streamline and increase the efficiency of this process. Moreover, software usually includes quality assurance procedures to control for and minimise data entry errors. Access rights need to be specified and implemented in the software to maintain security while allowing different types of users to access and manage the data. The choice of data management software may depend on the mode of assessment delivery (i.e. paper- or computer-based) and data entry/capture software, as well as budget, timeline, etc.:

*Manual data entry* requires personnel to manually enter assessment data into the software. As such, it requires a large amount of human resources and entries are more prone to error. Error can be controlled to a certain extent, through functions embedded in the software that define acceptable entries and automatically check that data adheres to those definitions.

*Optical character recognition or optical mark recognition* greatly reduces the time required for data entry by automatically converting the paper-based data into digital formats using scanning hardware and specialised software. While responses to multiple choice items in test and questionnaire/interview forms can usually be automatically converted, open constructed response items still require scoring through expert scorers. Expert scoring can also be done electronically using a specific scoring program where the images of the constructed response

items that were created during scanning are presented to the scorers who manually enter a score for each response in the scoring program.

*Digital data processing* as part of computer-based assessments is most efficient in terms of data processing, where participants' responses are recorded directly in digital format. However, this option entails considerable costs in the development of the delivery software, development of the items, and the purchase/provision of the devices. In addition, this option requires the highest technical know-how for the test administrators.

**Establish and train a data manager and data management team.** The data manager is the member of the assessment team who will be responsible for data capture, data cleaning, verification and validation, and the training of data processing personnel. Data managers should be thorough and experienced in managing larger sets of data. Recruitment of the data manager and team should start early to ensure data can be processed in a timely manner. Training in the use of software and in the relevant data capture and data cleaning protocols is essential for all data management staff.

**Prepare protocols for data capture, data cleaning, verification and validation.** Protocols establish rules for data processing. Before processing begins, documentation outlining the protocols should be prepared.

*Data capture protocols* include categorising missing data, monitoring the quality of the manual data entry or optical recognition, and backing up data.

*Data cleaning, verification, and validation protocols* include procedures for checking and correcting unique identifiers (IDs), duplicated or dropped records (e.g. in the process of merging two datasets); checks that the values entered are valid and within range; checks that there is consistency between variables that link separate data files; and checks of logic when the response to one variable depends on another. Some discrepancies may require checks using external sources of data to resolve (e.g. school records, census data, public records and other pre-existing databases that are separate from the current data collection process). Protocols should also be in place in the case where inconsistencies are unable to be resolved, for example, if a record needs to be removed from the database due to issues with the unique identifier.

**Prepare data backup protocols.** Both raw and processed data should be backed up, following a regular schedule. For example, raw data need to be backed up before exporting into another format. Data should also be backed up whenever data files are consolidated.

**Prepare data documentation and transfer protocols.** A thorough *documentation* of the database is essential to ensure its adequate use. Database documentation should include the data source, a description of the database content and structure (e.g. nature and order of cognitive and contextual data in the database, scales and indices, proficiency levels; information about recoded data; information about sample weights and replicate weights [see Key Area 9]), and the data codebook. If the database includes multiple datasets (e.g. for students, parents and schools), the documentation should also contain information on how these can be linked.

*Data transfer protocols* need to be prepared for different forms of data transfer/recipients (e.g. internally, externally to partners, or publicly), as each recipient may have different requirements. For example, de-identification might be required for external transfer or public release. For each transfer a systematic record should be kept.

# Key Area 12: Scaling cognitive and contextual data

## Objective

Cognitive and contextual data that is scaled using well-developed analytical tools in order to support a range of useful comparisons and to communicate information that is meaningful to a range of users.

***Cognitive and contextual data that is scaled using well-developed analytical tools…***

Assessment results are better interpreted and therefore more useful if they are reported on a scale sharing a common unit for measuring and interpreting results. The term 'scaling' refers to the process of converting raw data into numerical indicators of the scale (Berezner & Adams, 2017). Scaling can be considered primary data analysis, and scaling outputs are usually released as part of the final dataset (see Key Area 11). Large scale assessments generally use Item Response Theory (IRT) to scale raw data in the production of cognitive and contextual constructs.

***…in order to communicate information that is meaningful to a range of users…***

The process of scaling results in a numeric reporting scale that is useful for statistical analysis. To improve its usability, this information is commonly enhanced with substantive descriptions of skill and knowledge progression along the scale. These descriptions are generally derived from either outlining expected performance (e.g. according to curriculum standards), or observing performance as indicated by the test results (Turner & Adams, 2017). Providing qualitative descriptions of what participants know and can do helps a larger range of users to better understand, interpret and ultimately respond to the data.

***…and to support a range of useful comparisons.***

The comparisons made in a large-scale assessment will depend on the policy goals set out at the beginning of the assessment. These comparisons are commonly between sub-populations of interest, for example, between rural and urban participants, but they are also often across grade levels and over time. Linking different tests using IRT methodology allows for comparisons to be made across different tests, settings and over time.

## Achieving the objective

**Develop a data analysis plan.** All scaling (e.g. scaling of cognitive and contextual data, item analysis, description of the scales and sub-scales)) and data analysis (e.g. for the purpose of reporting, see Key Area 13) should be planned as early as possible during the development stage of an assessment program. Based on the initially defined policy goals and priorities (see Key Area 1), the analysis plan should describe the kind of data and analysis required to address these. Thus, the analysis plan provides important information for instrument development (see Key areas 4, 5, and 6) and design (see Key Area 8).

The analysis plan should also consider how many personnel will be involved in the analysis and its preparation, their specific roles, and the time available to conduct the analysis. Another important consideration is the analysis software to use, and which is best fitted to the analysis needs of the assessment and the capacity of personnel.

For the purpose of *scaling* the plan should specify the analytical model to be applied, the scales and sub-scales to be created, the analysis undertaken with the cognitive data, the description of the cognitive scales and sub-scales, and analysis of contextual data.

**Choose analytical model.** Decisions regarding the analytical models used will likely be made during the development of the analysis plan. Large scale assessments generally use Item IRT to scale raw data. Within IRT, there are a number of different approaches used. These approaches are described below along with the differences between Classical Test Theory and Item Response Theory.

- *Classical Test Theory vs Item Response Theory*. Classical Test Theory (CTT) focuses on estimating each participant's 'true score' and making inferences about his or her likely score on a test. Ability is usually described within the boundaries of 0% and 100% correct on a test. Although the item difficulties and participant results can be viewed alongside each other to aid interpretation, CTT is limited to comparing scores on the same test. There is little scope for generalising skills of participants at specific ability levels. Comparing performance over time is not possible unless the same tests are used each time. The major limitation of CTT is that the, observed scores are item/test dependent, and the item statistics are sample dependent. IRT was developed to address these main limitations of CTT.

  IRT models focus on estimating each participant's ability and making inferences about each participant's ability level on the construct (i.e. a latent trait such as intelligence, motivation or maths ability) that is being tested. Unlike CTT, a latent trait is measured on an infinite continuum (e.g. not only between 0 and 100), where the measurement unit is denoted as a logit. If a mathematics test is given, the IRT approach would try to estimate each participant's level on the latent trait of mathematics. The logit defines distances between differences in scores which can be easily interpreted. It also can link item scores to person scores. IRT offers more capacity than CTT for linking different tests and providing substantive interpretations to scores on a test. It helps in placing different tests on the same scale for comparison in time.
- *Rasch (one-parameter logistic) model vs Birnbaum (two- and three-parameter logistic) models.* These models are derived from IRT. The Rasch model is often referred to as the one parameter logistic model because it uses a single parameter to describe each item. Birnbaum models introduce additional parameters that describe additional features such as the strength of the relationship between the item and the construct (a discrimination parameter), or the probability of success on the item through random guessing (a guessing parameter). The Birnbaum models are therefore more general than the Rasch model and provide a better fit to the data that are collected—some argue that this better fit means they are a more valid representation of the data (Berezner & Adams, 2017). With a stricter definition of what constitutes measurement, others argue that the Rasch

model has stronger construct validity (Berezner & Adams, 2017). The Rasch model is also more stable over time.

**Identify scales and possible sub-scales.** The scales and sub-scales of a cognitive domain are usually defined and described during the assessment framework development. All items will subsequently be written according to the framework (see Key Area 5). As such, it is essential that information about which items correspond to which learning domains and sub-domains be provided to those conducting the scaling.

Information about which contextual items intend to measure the same characteristic (or construct) is equally essential in constructing contextual scales (or indices).

**Analyse cognitive data.** As IRT is most commonly used in large-scale assessments, the following activities refer to those followed when using IRT. While all analyses described below should be conducted for both the field trial and the main survey, the focus for each is different. For the field trial, the focus is on the selection of items for the main survey and ensuring that the items reflect the learning domain being assessed. For the main survey, the focus is on ensuring the reliability of results (Berezner & Adams, 2017).

*Calibration.* Scale calibration is the process of estimating the parameters of the model (e.g. item difficulty in Rasch) and placing these parameters on a uniform scale (Kolen & Brennan, 2004). It involves a number of processes which include: reviewing the model fit of the items to determine if it matches the data; assessing each item's fit, differential item functioning, and content, to determine if it is retained for the final scaling of responses; reviewing the test targeting to ensure that the difficulty of the items matches the ability of the participants; and reviewing any anchor/link items to ensure that anchor/link items are working or behaving well across different tests.

*Estimate participant ability*. Once the items are located on a single scale, participant scores can be computed. Participant ability estimates can be categorised into two main groups: point estimates and plausible values. Point estimates use a single value to estimate ability and are best for reporting on individual scores. Plausible values use a set of values (usually five) to estimate ability and are best for reporting at the group level (von Davier, Gonzalez, & Mislevy, 2009). Standard errors are also important in estimating ability as they help to determine the precision of the parameter estimates (see Key Area 13 for more detail).

*Linear transformation.* Linear transformation is the process of transforming scores in logits, to a chosen mean and standard deviation. This allows scores to be reported from a test (or several tests) on a readily understandable scale. This is particularly helpful when providing information related to content, norm or reference groups, and trends.

**Describe the cognitive scales.** IRT results in the mapping of test items and participant ability onto a single scale. Once this scale has been defined numerically, the skill and knowledge progression along the scale should be substantively described. This ensures that the information is accessible for a large range of users of the data. These descriptions are generally derived from either outlining expected performance (e.g. according to curriculum standards), or observing performance as indicated by the test results (Turner & Adams, 2017). Descriptions

should be revised and refined after the main survey. Developing the scale descriptions generally takes the following into consideration:

*Determine the number of levels*. The number of described performance levels will depend on the policy goals of the assessment and the amount of detail needed to describe performance. For example, benchmarks that indicate basic, proficient and advanced skills and knowledge may be sufficient for determining the quality of the education system, while on the other hand, more levels may be needed if the data is to be used by schools and teachers to target learning interventions.

*Define the levels on the scale*. As the scale is a continuum, there is no natural point where the levels on the scale should be assigned. It is therefore essential to define what it means statistically to be located within a level. Three principles can be applied to define each level (Turner & Adams, 2017): the first is to define how successful a participant at that level should be in answering the items in the level, for example, participants performing at the bottom of the level should answer 50 percent of the items correctly; the second is to ensure that the width of each level is similar; and the third is to determine the proportion of participants expected to answer each item in the level correctly.

*Summarise skills within each level*. During item development, the characteristics of each item (metadata) should be recorded (see Key Area 5). This information can then be used to develop a summary definition of the skills and knowledge that participants have when they perform at or below the particular level.

**Analyse contextual data.** The aim of collecting contextual data in large-scale assessments is to provide context to the performance data. Scaling of contextual data is usually a reflection of the participant's likelihood of agreeing or disagreeing with a particular statement (e.g. 'I like mathematics'), or of a composite index that aims to describe several factors of an underlying characteristic (e.g. socioeconomic status) using one score (Schulz & Lietz, 2017). The following analysis are typically undertaken with contextual data.

*Ensure items measure the underlying characteristic/construct*. Similar to cognitive items, analyses should be performed to review the dimensionality of contextual items. However, due to the categorical nature of most contextual items, different analyses can be performed in conjunction with IRT modelling. These include CTT, Exploratory Factor Analysis and Confirmatory Factor Analysis. IRT scaling plays a vital role in contextual construct evaluation and development. IRT modelling provides information not only on the performance of each contextual item in the scale and how the scales functions overall for measuring a construct, but also provides an elegant way of dealing with missing data.

*Scale contextual items.* While scaling is possible using CTT, IRT is preferable as it takes into account the fact that some items may be harder to agree with than others, or some factors may contribute more to the underlying characteristic/construct than others.

# Key Area 13: Analysing data

## Objective

Analytical results that are fully documented and reproducible, and that permit valid and useful inferences about the population(s) of interest.

***Analytical results…***

The term 'analytical results' here refers to analysis undertaken for the purpose of reporting and interpretation. In particular, analyses of relationships between variables (e.g. between achievement and geographic region) is of interest in policy-level reporting. Scaling (see Key Area 12) is a related area of data analysis the results of which are usually released in a public database.

***…that are fully documented and reproducible…***

All analyses should be accompanied by a description of the approaches used to analyse the data. This enables other users of the data to reproduce the results if they wish. The reproducibility of the results plays an important role in verifying their accuracy, and it turn, helps to ensure that statements made on the basis of the assessment results are also accurate.

***...and that permit valid and useful inferences about the population(s) of interest.***

The analysis should answer questions related to the assessment's policy goals, to ensure that it is consistent with the purpose of the assessment. This will help to strengthen its relevance for stakeholders and its usefulness in making policy and practice decisions. However, analyses are only useful if they are technically sound and appropriate for the data available. If results are based on analyses that are inappropriate, then their validity is compromised and they will not provide accurate information for education decision making.

## Achieving the objective

**Develop a data analysis plan.** The analysis to be undertaken for reporting are essentially planned together with the primary analysis (see Key area 12), and ideally during the development phase of an assessment program.

The analysis plan should describe the initially defined policy goals and issues and how these are going to be addressed with the analysis of the cognitive and contextual data in the database.

The analysis plan should also consider how the results will be reported, e.g. in tables, graphs, etc. After the data collection stage, the analysis plan should be refined. This ensures that the analyses can be carried out in a targeted manner.

**Assign sample weights.** In most situations, samples do not precisely represent the population and therefore the population estimates derived from them would be biased due to this misrepresentation. A correction technique, through the use of sampling weights (see Key Area 9), can be used to adjust the sample and reduce the bias in the population estimates. Any analysis should, therefore, always be weighted at any stage of the process, whether it is the primary or secondary data analysis.

**Calculate the standard error.** A standard error is the spread or variability of a sample statistic around its mean. In other words, it is a measure of the accuracy of a sample statistic as an estimate of an unknown population parameter. It is essential to report accurate and unbiased standard errors, as these estimates are used for calculating the statistical significance of analysis results. The statistical significance describes the probability that a sample statistic is likely to be a true reflection of the population, or just a result of sampling and measurement error.

To achieve unbiased standard errors from survey studies, the analysis must have accurate estimates of both sampling variance and measurement variance (OECD, 2012). Calculating the sampling variance depends on the way in which the sample was obtained (simple random sample or complex random sample—see Key Area 9). Calculating the measurement variance usually takes the plausible values outlined in Key Area 12, and calculates the variance among the plausible values. If the results are to be analysed across time (i.e. from previous administrations of the assessment) a linking error should be added as a third component of the standard error.

In a report of the results, each population estimate should be accompanied either by its confidence interval or standard error, along with the statistical significance of any comparisons.

**Analyse data.** The data collected in an assessment of learning, outcomes can be analysed and described in a number of ways. This includes frequency analysis, comparing mean scores, and comparing the variance in scores across groups. Since many educational policy questions concern the relationship between performance and other variables, a commonly used technique for educational assessment data is regression analysis. Regression analysis predicts an outcome variable (e.g. achievement) using one or more explanatory variables (e.g. gender, language, region). Depending on the analytical model or research question, different kinds of regression analyses can be performed.

**Analyse trends**. There are a number of issues to consider when conducting any trend analysis. Trends over time on any indicator require careful interpretation, and need to consider how contextual factors may have changed that impact the inferences that may be drawn from trend analyses, such as changes to the level of government funding provided for education after a change in government. In large-scale assessment research, it is also important to consider how comparable the definition of the trend indicator is for subgroups within the population. For example, it is important to consider the different types of home possessions that indicate wealth in urban and rural populations. Reliable horizontal trend measures depend on consistency over time in a) the comparability of the target population b) the data collection procedures, and c) the assessment framework. Trends can be computed directly without any precautions when the data collected at two different time points are 'linked' on a common scale (see Key Area 8 and Key Area 12).

# Key Area 14: Reporting and dissemination

## Objective

Appropriate products and approaches to reporting and dissemination that are tailored to the different stakeholder groups and promote appropriate and effective use of the assessment results by those groups.

***Appropriate products and approaches to reporting and dissemination…***

Appropriate reporting methods and dissemination strategies need to be developed to support stakeholders in understanding and making effective use of the assessment results. Assessment products could include various types of reports (e.g. summary reports, main reports, thematic reports), the public release of the database and press releases. In addition to reporting, other dissemination approaches could include workshops, conferences, media appearances, websites, etc. See Table 2 for a description of reporting and dissemination products and approaches.

***...that are tailored to different stakeholder groups…***

Assessment results may be used by a wide variety of stakeholders to inform discussions and debate around the results themselves and possible policy responses (Kellaghan, Greaney, & Murray, 2009). For the dissemination products and approaches to be effective, they need to take into consideration the various stakeholder groups who will be using the results. This includes considering the information needs of the target audience—teachers may aim to better understand students' learning strengths and weaknesses while policy makers may aim to identify under-resourced areas—the expected technical knowledge of the target audience, and the most effective communication method for the target audience. One important stakeholder group is the public, not only because many people are involved in education as parents or students, but also because assessments are often financed through government, hence public, funding and therefore under particular scrutiny. Therefore, the media—traditional media outlets as well as social media—have to be given special consideration as the main drivers of public perception.

***...and promote appropriate and effective use of assessment results by those groups.***

More often than not, assessment results describe problems, but they do not specify solutions (Kellaghan et al., 2009). Therefore, if learning assessments are to bring about change, stakeholders need to understand the meaning and relevance of the results, be able to use the results to identify appropriate actions, and be in a position and system where they are able to enact or support change. Reporting and dissemination should take these needs into account and aim, not only to inform stakeholders, but also to build their capacity in using the results for improving learning.

## Achieving the objective

**Identify different information needs of stakeholders.** Consultation throughout the assessment program with key stakeholders such as the steering committee (see Key Area 1), representatives from curriculum agencies, education sectors, teacher training institutions and unions, parent associations, journalists and others, can be used to establish what the relevant policy and practice issues are, and to gain a deeper understanding of stakeholders' requirements.

Consideration should also be given as to the unintended consequences of reporting results. For example, high stakes assessments tied to funding decisions can have unintended negative consequences such as resistance from schools to participate, a narrowing of the curriculum, teaching to the test (Braun, Kanjee, Bettinger, & Kremer, 2006) or cheating, all of which can lead to a corruption of data.

**Develop a dissemination strategy.** A dissemination strategy should be developed early in an assessment program, so that dissemination methods can be planned to occur throughout the program to maximise stakeholder engagement. Flexibility is also needed as new stakeholders are identified, as resources become available, as evidence is gathered about what dissemination methods are most effective, and as other possible policy implications of the assessment results become apparent. Ensuring that dissemination products can be easily and broadly accessed by a variety of stakeholders will help to increase the likelihood that results will be considered and used by a variety of stakeholders in decisions about education policy and practice.

There are benefits and drawbacks to all dissemination methods, so anticipating likely issues or questions and preparing for these is essential. The use of simple language and clear and consistent messages in all dissemination methods will aid understanding, and technical information should be available to substantiate all statements made.

**Develop dissemination products.** While a detailed written report is often one of the requirements of assessment funding bodies, a mix of dissemination methods and products probably addresses best the information needs of different stakeholders. Developing reports is generally an expensive and time-consuming process and as such, specifications should be agreed upon by the assessment agency and stakeholders. Different types of reports are described in Table 2 along with other dissemination methods. Some issues that affect all dissemination products are described below:

- *Assessment limitations*. Every assessment has limitations regarding what can be analysed and the inferences that can be made. Reporting should make clear these limitations to ensure results are reported accurately and used appropriately by stakeholders. In particular, the possibility of over-simplistic interpretations of the assessment results by stakeholders should be addressed and discouraged in reporting, e.g. league tables.
- *Summary of findings.* In order to draw readers' attention to the most important information, a summary of findings or an executive summary section should be included at the start of a report. In addition, reiterating the core messages increases the likelihood that audiences will engage with the reporting products. As such, it can be helpful to also include a short bullet-point list of the key information at the start of each chapter.
- *Implications and recommendations*. In concluding a report, some reports may build upon the summary of findings and highlight the relevance of key results for broader policy, practice, and research through the inclusion of implications or recommendations. Implications may include general inferences suggested by the assessment results, while recommendations may refer to more specific suggestions. Dependent upon the key findings and availability of external data, implications and recommendations may be

based on the assessment data alone, or may also draw on findings from other assessments, evaluations or research.

- When developing implications and recommendations, it is important to consider stakeholders' expectations of the types of implications and recommendations, the assessment team's ability to formulate useful implications or recommendations based on the assessment data, the opportunities stakeholders have had to be involved in discussing the assessment findings (implications or recommendations which have been discussed with stakeholders who have in-depth knowledge of the sector are likely to be more robust), and the technical quality of the learning assessment, and therefore, the validity of the implications or recommendations.
- *Longevity*: Consider scheduling the release of various dissemination products over an extended period of time. The purpose of such an approach is not only to maintain interest and momentum in the assessment but also to instil a view that education is an enterprise where growth and change needs monitoring and sustaining over an extended period.

**Monitor how assessment data are used over time.** When possible, it is informative for assessment agencies to keep track of the different ways that assessment data are and are not utilised by various stakeholders within the education system. This will help assessment agencies evaluate their dissemination and reporting strategies, and have a better understanding of how to target the policy and information needs for different stakeholders within the education system for future assessment cycles.

**Table 2: Dissemination methods**

| Dissemination method | Description and purpose | Main audiences and level of technical detail |
|---|---|---|
| Executive summary report | Provides a summary of the key findings and policy-related messages that emerge from the first analyses of the data. Can sustain interest in the assessment and drive the policy agenda in the period between data collection and publication of the main report. | All stakeholders, researchers, educational practitioners, media and the public.<br>Level of technical detail: low |
| Main report | Provides an overview of all aspects of the assessment so that a variety of stakeholders can understand the purpose, approach taken, results and implications. Almost all assessment programs have a main report. | All stakeholders, researchers, educational practitioners, media and the public<br>Level of technical detail: medium |
| Summary reports, pamphlets | A summary of the important points from the main report to provide a fast way for stakeholders to learn about the most important assessment results. The summary reports can vary in length. | May be produced for a variety of stakeholders, including teachers, policymakers, the general public or key interest groups<br>Level of technical detail: low |
| Technical report | Provides detailed information about the assessment processes and data collected to judge the quality of the assessment and to inform the interpretation of results. It also serves a record of activities which can inform future assessment phases. If a separate technical report is produced, some technical details can be left out of the main report, making it more accessible. | Key stakeholders and researchers<br>Level of technical detail: high |
| Assessment framework report | Provides details about the assessment framework that guided the development of the assessment – the cognitive learning domains as well as the contextual information collected. The framework usually includes a definition of the cognitive learning domains and explains all aspects that are measured in detail and how, including example items. The framework may also outline how the results of the assessment will be reported (e.g. described performance scales).<br>A summary of the assessment framework may be included in the main report; however, the full assessment framework may be published as a separate report, either before, during or after assessment implementation. | Key stakeholders, researchers, educational practitioners and the public<br>Level of technical detail: medium to high |
| Thematic reports | Reports that provide more detailed information than the main report around a particular topic of interest (e.g. a report on differences in achievement patterns between girls and boys). Producing thematic reports can help raise awareness about a particular priority area. | Particular stakeholder groups, researchers<br>Level of technical detail: medium to high |

| Dissemination method | Description and purpose | Main audiences and level of technical detail |
|---|---|---|
| Policy briefings | Short briefings provide a summary of the main information and possible implications. These messages can be communicated concisely to decision-makers who do not have time to read a full report. These decision-makers can use this information to identify possible next steps. Can be written or delivered by presentation. | Ministers and policymakers.<br>Level of technical detail: low to medium |
| Media reports | Can include newspaper articles, radio or television reports, blogs, videos and press conferences. Allows information to be spread to a wider audience in an accessible way. However, care must be taken as the media may greatly simplify assessment results or focus only on more controversial results (e.g. league tables). | The public<br>Level of technical details: low |
| Press releases | Short written statements provided to the media that succinctly communicate factual information about the assessment, what the program assesses and how it is conducted, and key findings from the assessment that are important for the wider public to understand. Encourages more accurate and reliable dissemination of results through the media to reach a wider audience in an accessible way. Allows for more control in what is reported by the media, to support the appropriate use of results for informing policy and practice. In addition, is a cost effective dissemination strategy. | The public<br>Level of technical details: low |
| Assessment database | Assessment data can be made publicly available or available to certain stakeholders/ organisations that have been granted access. These can be used to investigate particular areas of interest. Usually requires training in the appropriate use and analysis of data. | Particular stakeholder groups such as government officials, researchers and organisations<br>Level of technical details: high |
| Conferences and workshops | Involves the discussion and presentation of the assessment to stakeholders. Workshops generally involve a smaller number of people and are more participatory than conferences.<br>Provides an opportunity to gather feedback from stakeholders and to discuss possible policy implications. | Particular interest groups, such as teacher trainers or curriculum developers, Particular stakeholder groups, researchers and organisations<br>Level of technical details: low- medium |
| Websites | A webpage for the assessment program may contain links to different dissemination outputs. For example, reports, press releases and the assessment database may all be found on a webpage. This enhances accessibility to different dissemination products. It may also contain an interactive display of the assessment database where users input the information they require, and receive an output that is identical to that in the main report (Cresswell, 2017a). | All stakeholders, researchers, educational practitioners, media and the public.<br>Level of technical details: low- medium |
| Blogs and social media | Blogs and social media disseminate assessment results and other assessment information in small packages that are easily accessible by a wide-ranging audience. Blogs and social media also encourage feedback from the public, which can provide a direct link between the assessment agency/ministry and public discourse. | All stakeholders, researchers, educational practitioners, media and the public.<br>Level of technical details: low |

| Dissemination method | Description and purpose | Main audiences and level of technical detail |
|---|---|---|
| Sample items and contextual instruments | Some items can be released to the public to provide a better understanding of what the assessment entails. Chosen items should be of high quality as they will represent the assessment. A possible source is field trial items that were suitable but not used in the main survey due to there being too many items of that format or difficulty level. A large proportion of the items should remain secure so they can be used again in the future (Cresswell, 2017a). Contextual instruments on the other hand are usually not secure and could be released in their entirety.<br>Sample items and contextual instruments are usually accompanied with information about the skills that item is assessing, and its relation to the framework. | All stakeholders, researchers, educational practitioners, media and the public.<br>Level of technical details: medium |
| Manuals | Manuals, such as sampling, data management, test administration, translations, etc., can be released to the public to provide a better understanding of what the assessment entails. Manuals might also be developed specifically for policy makers and researchers to help them use the database. | All stakeholders, researchers, educational practitioners, media and the public.<br>Level of technical details: high |
| Analytical services | The assessment agency may offer analysis services to the public. Extra analysis will likely cover that not addressed in the final set of assessment reports. This ensures that the data is widely used, and not dependent on stakeholders' high level of technical expertise. | All stakeholders, researchers, educational practitioners, media and the public.<br>Level of technical details: low- medium |

# Appendix

## Mapping GP-LA to UN Fundamental Principles

The GP-LA interacts with the Fundamental Principles of Official Statistics on two levels: the GP-LA itself aims to adhere to the principles; and the GP-LA operationalises the principles in the context of generating statistics on learning outcomes through 14 key areas of a robust learning assessment. Table 3 maps the interactions between the GP-LA and the UN Fundamental Principles of Official Statistics, with relevant key areas described previously highlighted using the prefix 'KA'.(United Nations General Assembly resolution 68/261; United Nations Statistics Division, 2015).

**Table 3: Relationship between the UN Fundamental Principles of Official Statistics and Good Practice in Learning Assessment**

| UN Fundamental Principles of Official Statistics | How the GP-LA adheres to the principles | How the GP-LA operationalises the principles |
|---|---|---|
| Principle 1: 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. | Relevant to all agencies involved in generating data on learning assessment<br>Produced in a neutral and unbiased manner with attention paid to the diverse range of contexts within which learning assessments are used<br>Accessible and available for all | Encourages consulting stakeholders (KA1, KA4, KA5, KA6)<br>Encourages the use of transparent processes for hiring staff and establishing expert committees (KA2, KA3, KA4, KA5, KA6, KA7)<br>Outlines considerations for reporting results objectively and in response to policy goals (KA1, KA14)<br>Outlines considerations for making data available to all and considering all stakeholders when reporting results (KA11, KA14) |
| Principle 2: 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. | Developed by experts in learning assessment<br>Methodologies outlined in the GP-LA are based on established scientific principles<br>To be used to guide other aspects of SDG 4 reporting including the Data Quality Assessment Framework | Encourages hiring professionals and developing capacity (KA2, KA5, KA6, KA7)<br>Encourages defining technical standards (KA3) and monitoring their adherence (KA5, KA6, KA7, KA10, KA13)<br>Outlines scientific principles for learning assessment (KA8, KA9, KA10, KA11, KA12, KA13)<br>Highlights ethical considerations in terms of fairness, inclusiveness, and confidentiality. |

| UN Fundamental Principles of Official Statistics | How the GP-LA adheres to the principles | How the GP-LA operationalises the principles |
|---|---|---|
| *Principle 3*: 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. | Collaboration with other experts in its development<br>Clear definitions of quality and how to achieve it in learning assessment<br>Released publicly | Encourages involving stakeholders through committees (KA1, KA2, KA3, KA4, KA5, KA6)<br>Encourages hiring professionals (KA2)<br>Encourages transparency through releasing methodological reports (KA14) and developing project websites (KA1) |
| *Principle 4:* The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics | Ensures that processes are not manipulated to produce specific outcomes by openly stating international standards for learning assessment | Encourages involving stakeholders in decisions with the aim to improve ownership of the results and their honest interpretation (KA1)<br>Encourages the development of educational material for key user groups (KA14) |
| *Principle 5*: 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 | Directs users to other sources of information and provides examples of good practice | Encourages avoiding the collection of data already available from somewhere else (KA6)<br>Outlines considerations for developing an assessment framework (KA4) and reliability measures (KA3) to facilitate use by other agencies<br>Outlines considerations for managing staff and contractors (KA2) |
| *Principle 6*: 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. | Maintains a focus on the use of aggregated data for describing learning outcomes | Encourages developing a staff and contractor confidentiality policy (KA2)<br>Outlines considerations for developing mechanisms for maintaining data privacy (KA10, KA11, KA14) |
| *Principle 7:* The laws, regulations and measures under which the statistical systems operate are to be made public. | Operates as part of the UIS mandate to monitor SDG 4 indicators on learning outcomes<br>Operates under the SDG 4 quality assurance framework<br>Publicly released to ensure transparency of processes | Encourages collaboration with stakeholders to ensure accountability and transparency of the learning assessment system (KA1) |
| *Principle 8:* Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system. | Developed in collaboration with agencies working with national governments | Encourages consulting stakeholders to ensure relevant statistics (KA1, KA3, KA4, KA5, KA6), unduplicated data (KA6) and the exchange of technical knowledge (KA3)<br>Acknowledges that assessment teams may be made up of staff from various agencies and/or contractors (KA2) |

| UN Fundamental Principles of Official Statistics | How the GP-LA adheres to the principles | How the GP-LA operationalises the principles |
|---|---|---|
| *Principle 9:* 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. | Adheres to the UN Fundamental Principles of Official Statistics | Articulates international standards of learning assessment |
| *Principle 10:* Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries. | Developed in collaboration with diverse learning assessment experts | Shares good practice internationally, develops a common definition of terms, and uses diverse examples to illustrate good practice around the world<br>Helps to identify capacity development needs |

*Source for UN Principles: (United Nations Statistics Division, 2015)*

# Glossary

| | |
|---|---|
| Access rights | Specific permission or authority granted to different types of users to access and/or manage a database. |
| Administration materials | Manuals relating to the administration of the tests and contextual instruments (otherwise known as field guidelines or field operations manuals) as well as important supporting documents such as student attendance forms (sometimes referred to as student tracking forms). |
| Assessment agency | The body tasked with the organisation of the assessment. It could be a standalone agency, or a team within an existing organisation like a university or the ministry of education. |
| Assessment design | The implementation plan for the whole assessment, including its purpose, the target population, the content to be tested, testing cycles, etc. |
| Assessment materials | Test forms, questionnaires, interviews, observation forms |
| Benchmark | A standard set as part of the assessment program (e.g. performance levels) or from outside the assessment program (e.g. SDG 4 learning outcome targets) against which to assess performance on the test. |
| Bias | A systematic distortion of results that is based on factors unrelated to ability. |
| Blueprint | A description of how the test will be constructed, including the details of the proportion of items that will assess different learning domains and skills and the response formats. Is sometimes referred to as a table of specifications. |
| Categorical response format | Where participants choose one or more response options from a list with no specific order. For example, girl/boy, urban/rural, brick/canvas/tin. |
| Census-based assessment | An assessment that is delivered to all people in a population. For example, an assessment of all Grade 3 students in a country. |
| Change log | A document that records the changes applied to a dataset, datafile, or generally any file that is being edited. |
| Cloud access | The ‘cloud’ is a term to describe a networked set of data centres. Among other things, access to the cloud allows the user to store and share large datasets. |
| Cloud storage | Digital storage space that is located in remote computer servers. Examples would be DropBox, OneDrive, Google Drive and other commercial data backup services. |
| Cluster (test and questionnaire design) | A small group of test/questionnaire items that are grouped together and treated as a block during test construction. |
| Cluster (sampling) | A sampling technique used when ‘natural’ but relatively homogeneous (similar) groupings are evident in a population of interest. |
| Codebook | A documentation of characteristics of the item that are needed at the time of data capture and analysis. This information includes a unique item identifier, the learning domain or subject that the item is measuring, and the correct answer. |

| | |
|---|---|
| Cognitive instrument | A set of items used to collect information about what the participant knows, understands and can do in a particular learning domain, or domains. |
| Cognitive laboratories | See piloting. |
| Cognitive skills | Skills, sometimes called 'processes', 'cognitive domains' or 'aspects', are the ways of thinking, or intellectual approaches, that develop as individuals become increasingly proficient in a learning domain. |
| Common reporting scales | A set of global scales of progress in learning used to situate and compare national assessment results at the global level. |
| Complex random sample | A sampling methodology where not all members of the target population have an equal probability of being selected. This can occur, for example, through clustering students in schools, or dividing (or 'stratifying') the population into regions. |
| Confidence interval | An interval that specifies a range of values for a parameter estimate, based on a predefined confidence level, and calculated from one sample of the population. The confidence level (usually 95%) for an interval indicates the proportion of intervals, computed from all possible samples, that includes the true value of the parameter being estimated. |
| Constructed response item | An item for which the student constructs, or generates, a response to the question. |
| Contextual framework | A formal documentation, often within the assessment framework, of why and how characteristics of the test-taker or the test-taker's environment are to be measured. See Key Area 4. |
| Contextual factors | Characteristics of the test-taker or the test-taker's environment that may have an influence on his or her educational outcomes. For example, the presence of a library in the school, or the participation of teachers in professional development activities may be correlated with assessment results. |
| Contextual information | Data collected through questionnaires/interviews/observations on a range of topics that are useful to policy and in understanding the test results in context. |
| Contextual instruments | A set of items used to collect information about the personal characteristics, background, attitudes and values of participants in their contexts (e.g. home, classroom, school). |
| | |
| Correlation | Indication of a relationship between two phenomena/variables. |
| Cross-sectional | An assessment where data are collected from individuals at a single point in time. While some assessment designs may collect data from, for example, a student cohort as they progress through school, that data is not tied to specific individuals. |
| Cycle (assessment) | All activities related to a single main survey assessment administration within a program with repeated administrations designed to assess learning over time. |

| | |
|---|---|
| Data capture | The act of recording test and contextual responses and other participant information into a database. |
| Data cleaning | The process of identifying discrepancies and errors in the database and correcting or removing them. This process includes verification and validation of the data. |
| Data collection | The process of gathering data—in the case of large-scale assessments, the process of administering tests and contextual instruments to participants. |
| Desired target population | The population to which inferences from the survey outcomes will be made. |
| Defined target population | The desired target population minus certain elements that are excluded due to practical difficulties. |
| Described performance levels (or scales) | In order to substantively describe the scale, it is divided into levels as it would not be practical to describe every score. The described performance level therefore describes the skills needed for participants to achieve a score at (or below) that level (Turner & Adams, 2017). |
| Differential item functioning | When the probability of answering an item correctly depends on the sub-population the respondent belongs to rather than her/his ability level. |
| Distractors | The incorrect options provided in a multiple-choice item. |
| Evaluation of Data Collection | The Evaluation of Data Collection is a process to be established to evaluate the technical rigour of the data collection process as part of SDG 4 reporting. The core of this evaluation are the methods and products from the data collection, for example the database and accompanying documentation such as technical reports, operational manuals, or results reports. |
| Field trial | Administration of items under test conditions, used to test the items' validity and the administration procedures. Occurs before the main survey and uses a sample as similar as possible to the target population. |
| Fit statistics | Indicators of model fit for both person data (i.e. a participant's response pattern) and item data (i.e. the pattern of responses to an item). |
| Form (test and questionnaire) | The group of test and/or questionnaire items that is presented to each participant. There may be just one group of items for all participants (i.e one form), or participants may receive one of several different groups of items (i.e. one of several forms). |
| Frequency analysis | A basic level of statistical analysis that describes the number of responses for each response category or score. For example, the total number of girls and the total number of boys. |
| Generalisability | The ability to make accurate generalisations about the whole population based on a sample of participants from within the population. |
| Historical linking | The linking of items between tests at the same grade level across different times/cycles. Can be used to estimate change over time. |
| Horizontal linking | The linking of items between tests at the same grade/age level in a single test administration to allow more items to be used than can be administered to a single individual. This allows more of the curriculum or learning domain to be used in the test. |

| | |
|---|---|
| Index (pl. indices) | A scaled indicator of a measure that is composed of several values or other measures. For example, a socioeconomic index might be composed of income, health factors, education level and other components. |
| Internal consistency | Internal consistency as a type of reliability estimate assumes that the test is unidimensional, or measuring a single construct. |
| Items | The questions or tasks used in an assessment. |
| Item descriptor | Description of what the item intends to measure. |
| Item difficulty | The difficulty of an item as hypothesised by test developers and confirmed by statistics. |
| Item discrimination | The ability for an item to group participants of different abilities. For example, participants who perform well overall on a test should also have a high chance of answering a particular item correctly. |
| Item pool | The total set of cognitive or contextual items written for an assessment. |
| Item statistics | The data used to assess whether items are functioning as they should (e.g. percentage of participants who correctly answered the item and average ability of participants who correctly answered the item). |
| Latent trait | A trait that is not directly observable, such as maths ability. Also called latent construct, it needs to be derived from a set of observed or indicator variables. |
| Learning domain | The area of learning that is the focus of an assessment. This may be a curriculum area (e.g. mathematics or science), or more generic areas of learning (e.g. reading, writing or problem-solving). |
| Linking items | Items which are common between tests and used for horizontal/vertical/historical linking. |
| Logit | Log odd units. This unit is based on the logarithm of odds ratio of an event. The odds ratio is the probability *for* an event divided by the probability *against* an event. Logits have a mean of 0 and standard deviation of 1. |
| Main survey | The final data collection stage where the data obtained is analysed with the aim of making generalisations and inferences about the desired target population. |
| Mean | The arithmetic average. |
| Measurement priorities | The specific statistical objectives addressed by the assessment. For example, to compare girls and boys, and rural and urban students, to identify those who are disadvantaged. |
| Metadata | A record of all the information related to an item, including the item code, the learning domain and skills the item is assessing, the estimated difficulty level and the item descriptor. |
| Mode of delivery | The way in which a test and/or contextual instrument is presented to the participants to complete. This could be in paper booklets, on computers, on tablets, in the form of an interview, etc. |

| | |
|---|---|
| Model fit | How well the overall distribution of the observed data (the data collected from participants) reflects the expected distribution according to the measurement model being used to analyse the data. |
| Multiple-choice item | An item that presents several options as answers, from which the participant selects one. |
| Operation plan | A comprehensive list of activities and resources and the timelines to which they need to be completed. |
| Outcome variable | In regression analysis, the outcome variable is the variable of interest which changes (or varies) depending on the predictor variables. |
| Outsourcing | Contracting an individual or agency located outside of the assessment agency to perform specific tasks. |
| Panelling | A process where a group of test developers (including the test developers who drafted the items) review and evaluate the draft items, looking for ways to make improvements. |
| Parameter | A characteristic that defines a population, such as its variability or its average. A characteristic that defines a sample is called a statistic. |
| Participant attendance form | A document used to collect student-level information such as identification variables, test form assignment and participation status for each test session. |
| Piloting | In assessment, piloting is also known as cognitive laboratories or cognitive interviews. These involve settings where participants are observed and studied in detail to investigate the thinking processes that they employ when performing assessment tasks. |
| Plausible values | A set of values drawn randomly from the marginal posterior distribution of scores that is used to represent performance in large-scale, sample-based assessments. |
| Point estimates | Estimates of parameters that relate to a single value of the corresponding statistic, often referred to as the 'best guess of a parameter'. For example, referring to one member of a population—a student score. |
| Policy goals | The overall purpose of the assessment. For example, to evaluate the equity of the education system. |
| Population | See 'target population' |
| Population distribution | The arrangement of people within the population. For example, the numbers of females and males, people within certain age groups, people with a disability, etc. |
| Predictor variables | In regression analysis, predictors are variables that are used in the regression model as explaining the outcome variable. |
| Processed data | Raw data that has been processed in preparation for analysis. For example, this could be student responses that have been scored and converted into values for correct or incorrect. |
| Psychometrics | Theory and methods of measuring psychological traits, such as mathematical ability or motivation to read. |

| | |
|---|---|
| Questionnaire | See contextual instruments. |
| | |
| Raw data | Data that comes directly from the source of the data and have not been processed in any way. These could be student responses on a test such as actual choices in a multiple-choice item, or actual words written in a short-response type of test. |
| Reliability | The consistency and accuracy of test and contextual measures and results over replications of the testing procedure (American Educational Research Association et al., 2014). |
| Reporting variable | Contextual factors that have been identified as important in accounting for the variance in performance across the target population, with the aim of discussing the outcomes in results reports. An example of a reporting variable could be gender, geographic location, or socioeconomic status. |
| Response formats | The ways in which students need to respond to the items (e.g. multiple choice, constructed-response). |
| Response rates | The number of sampling units that participated in the assessment (e.g. households, individuals) divided by the number of units that were sampled, expressed as a percentage. |
| Sample coverage | The percent of the desired target population that is covered by the defined target population. |
| Sample frame | A list of all the sampling units for a sample survey. For example, for an educational assessment survey of Grade 8 students in a particular state, the sample frame will contain a list of all the schools in the state containing Grade 8 students, along with an estimate of the number of Grade 8 students in the school. |
| Sampling weights | A statistical procedure used to correct for imperfections associated with the sample that might lead to bias and other departures between the sample and the population of interest. |
| Scale | A numeric or substantive description of progress in learning. |
| Scorers | Scorers, raters or coders are the people responsible for scoring the participant responses to items or tasks. |
| Scoring | The process of classifying responses and allocating (usually numerical) codes to represent the various categories of response. |
| Scoring guide | The description of the scoring categories that are used to categorise and score a participant's answer. |
| Selected response | An item response format where participants choose an answer from a given set. |
| Separation index | The 'ratio of the unbiased estimate of the sample standard deviation to the root mean square measurement error of the sample' (Wright & Stone, 1999, p. 162). This ratio indicates how large the model error variance is in proportion to the true variance. |

| | |
|---|---|
| Simple random sample | A method of sampling where every member of a population has an equal chance of being selected. |
| Skills | The ways of thinking, or intellectual approaches, that develop as individuals become increasingly proficient in a learning domain (sometimes called 'processes', 'cognitive domains' or 'aspects'. |
| Source language | The language used in the source version of the units. The source version is the version of the units upon which all translations in the assessment are based. |
| Standard deviation | A numerical measure of how the data values are dispersed around the mean. |
| Stem | The part of the item that contains the question or task (e.g. in a multiple-choice item, the part that introduces the options). |
| Stimulus material | The prompt or context on which one or more items is based. For example, in a reading test, the stimulus is often a prose text made up of one or more paragraphs. In a mathematics test, the stimulus may be a diagram or a graph. |
| Strands | The content categories that are to be included in the test which are specific to the learning domain. For example, in mathematics, typical strands are number, space, measurement and statistics. |
| Sub-population | Groups of people within the larger population who are separated into mutually exclusive categories according to a particular characteristic. |
| Sub-scale | A numeric or substantive description of progress in learning within a particular sub-domain or strand. |
| Substitute unit | A unit used to replace a sampled unit should the sampled unit be unable to participate. The substitute unit closely matches the sampled unit on pre-defined criteria. |
| Target language | The language used in the target version. The target version is the translated version of the test or contextual instrument. |
| Target population | A particular group of people that the assessment is attempting to describe or measure outcomes for. For example, an assessment may aim to measure reading ability of Grade 6 students in government schools in a particular region. This group of people is referred to as the target population. |
| Test targeting | In the context of test design, test targeting refers to the process in which item difficulties are matched with the ability levels of the target population. |
| Translation guidelines | Advice to translators on managing common translation issues in educational assessments, or on translating specific parts of text. |
| Trends | The change in assessment results over time. |
| Unit (sampling) | An individual element of the population used in sampling. For example, in cluster sampling, the first sampled unit may be the village, then within each sampled village, the second sampled unit may be the household. |

| | |
|---|---|
| Validity | The extent to which the assessment instruments measure what they claim to be measuring for a specified population, and the extent to which interpretations made from the data analysis are correct and appropriate for the proposed use of the data (American Educational Research Association et al., 2014). |
| Variance | A numerical measure of how the data values are dispersed around the mean. |
| Vertical linking | The linking between tests administered to different grade levels or age groups at the same time, achieved by using common items. Can be used to estimate growth between grade levels or age groups (e.g. comparing Grade 3 mathematics performance in 2015 to Grade 6 mathematics performance in 2015). |

# Bibliography

ACER. (2005). *PISA quality monitor manual*. Melbourne: ACER.

ACER. (2011). PISA MS 2012: CBA systems diagnostic manual (v 2.0). Melbourne: ACER.

ACER. (2013). *Monitoring educational development in Afghanistan: Project manager's manual, version 2*. Melbourne: ACER.

American Educational Research Association. (2011). *Code of Ethics*.

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.

Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for a national assessment of educational achievement*. Washington, DC: World Bank.

ASER Centre. (2014). Instruction booklet. New Delhi: ASER Centre.

Berezner, A., & Adams, R. (2017). Why large scale assessments use scaling and Item Response Theory. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.

Bordens, S. K., & Abbott, B. B. (1996). *Research design and methods: A process approach (3rd ed.)*. California: Mayfield Publishing Company.

Braun, H., Kanjee, A., Bettinger, E., & Kremer, M. (2006). *Improving education through assessment, innovation, and evaluation*. Cambridge, MA: American Academy of the Arts and Sciences.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioural research: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Cresswell, J. (2017a). Dissemination and reporting. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.

Cresswell, J. (2017b). Quality assurance. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.

Dept, S., Ferrari, A., & Halleux, B. (2017). Translation and cultural appropriateness of survey material in large-scale assessments. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.

Floyd, J., & Fowler, J. (2009). *Survey research methods (4th ed.)*. Thousand Oaks, CA: Sage.

Foy, P. (2000). Sampling weights. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report: IEA's repeat of the Third Mathematics and Science Study at the eighth grade*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. London and New York: Taylor and Francis.

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

Johnson, R. B., & Christensen, L. (2014). *Educational research: quantitative, qualitative and mixed approaches* (5 ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Kellaghan, T., Greaney, V., & Murray, T. S. (2009). *Using the results of a national assessment of educational achievement*. Washington, DC: World Bank.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer-Verlag.
Lietz, P. (2017). Design, development and implementation of contextual questionnaires in large-scale assessments. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.
Mendelovits, J. (2017). Test development. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.
Murphy, M., & Schulz, W. (2006). *Sampling for national surveys in education*. Melbourne: ACER.
NCERT. (2015). *Large-scale learning assessments: a handbook for the Indian context*. New Delhi: RMSA-TCA.
OECD. (2005a). *PISA 2003 technical report*. Paris: OECD.
OECD. (2005b). *School sampling preparation manual: PISA 2006 main study*. Paris: OECD.
OECD. (2012). *PISA 2009 technical report*. Paris: OECD.
Organisation for Economic Cooperation and Development. (2015). *PISA 2015 Technical Standards.*: OECD.
PEEC, & ACER. (2016). *Handbook on the development of a national assessment program for the Kingdom of Saudi Arabia*.
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.
Rust, K., Krawchuk, S., & Monseur, C. (2017). Sample design, weighting and calculation of sampling variance. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.
Schulz, W., & Lietz, P. (2017). Scaling of questionnaire data in international large-scale assessments. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.
Shiel, G., & Cartwright, F. (2015). *Analyzing data from national assessment of educational achievement*. Washington, DC: World Bank.
Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and applied multilevel analysis (2nd ed.)*. London: Sage.
**Thompson, S. K.** (1992). *Sampling.* John Wiley & Sons, Inc., New York.
TIMSS. (1994). *TIMSS main study manuals*. Hamburg: IEA.
Turner, R., & Adams, R. (2017). Describing learning growth. In P. Lietz, J. Cresswell, R. Adams, & K. Rust (Eds.), *Implementation of large-scale assessments in education*. New York: Wiley.
UNESCO Institute for Statistics. (2016). *Laying the foundation to measure Sustainable Development Goal 4*. Montreal: UNESCO Institute for Statistics.
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
von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series, 2*, 9-36.

Wright, B., & Stone, M. (1999). *Measurement essentials*. Wilmington, Delaware: Wide Range, Inc.