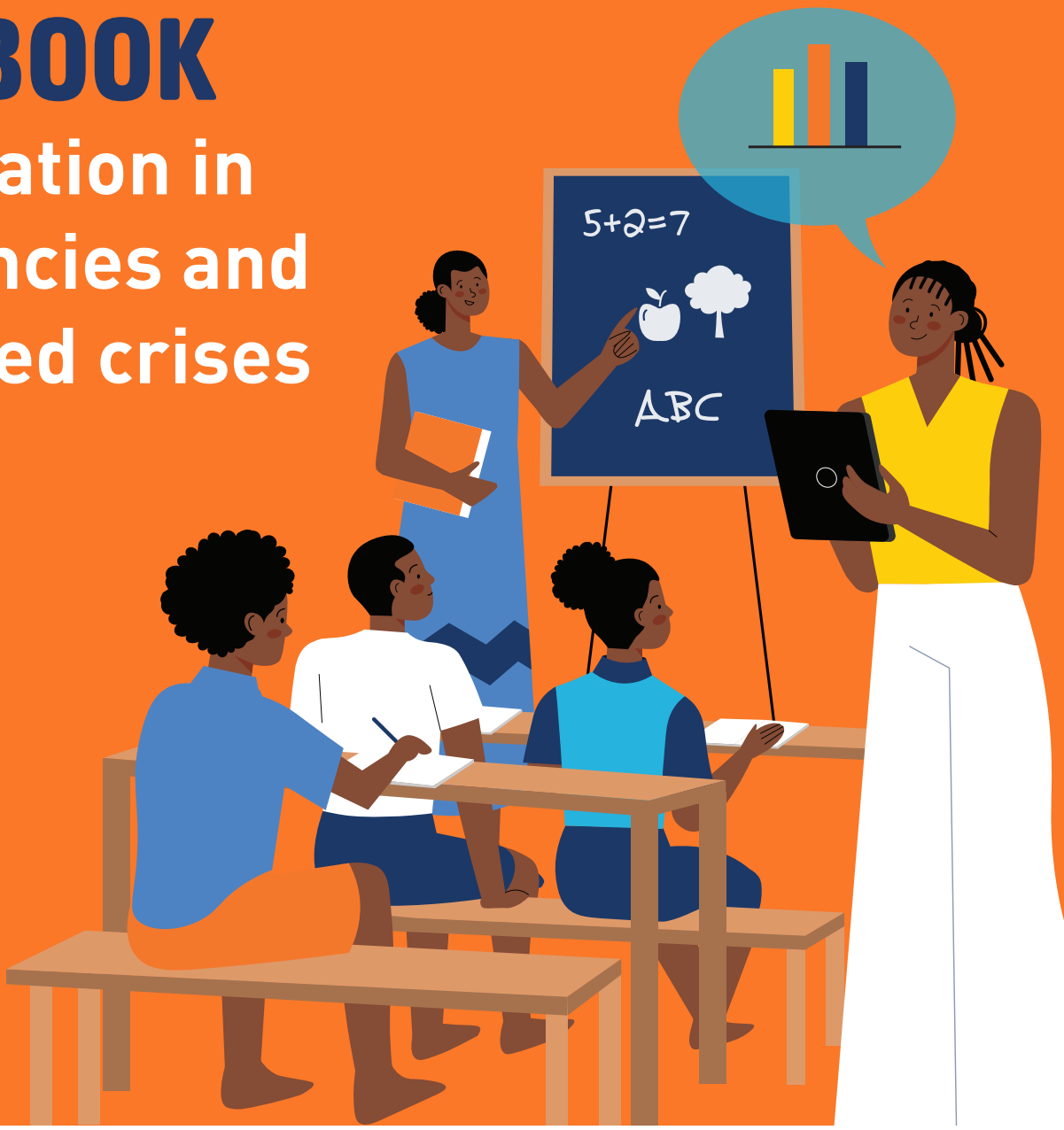
HOLISTIC LEARNING OUTCOMES MEASUREMENT HANDBOOK

for education in emergencies and protracted crises



CAMBRIDGE EDUCATION

Oxford MeasurEd

EDUCATION CANNOT WAIT

About this publication

This assessment Handbook was made possible through the collective efforts and support of several stakeholders and is part of Education Cannot Wait's Holistic Learning Outcome Measurement (HLOM) Initiative. The global technical partner (GTP) – Mott MacDonald Cambridge Education and Oxford MeasurEd – managed the HLOM, providing the blueprint for measurement work and technical assistance to grantees and national assessment partners. The GTP also leads the global report writing on HLOM findings across all participating countries. The authors of this Handbook are Rachel Outhred, Peter-Sam Hill, Fergal Turner and Paulina Valenzuela.

This Handbook was made possible under the coordination and technical guidance of Maurits Spoelder. The resource also benefited from the valuable feedback and guidance of Nicholas Santcross and Christian Stoff.

This publication was made possible thanks to the generous support of Porticus and Education Cannot Wait.

Copyediting: Jelena Borak
Layout and design: Svenja Greenwood

For any questions or feedback, contact Maurits Spoelder, mspoelder@un-ecw.org.

The views expressed in this publication are those of the author(s) and do not necessarily represent those of the United Nations, donor agencies or United Nations Member States.

© Education Cannot Wait, September 2025

HOLISTIC LEARNING OUTCOMES MEASUREMENT HANDBOOK

for education in emergencies and protracted crises

Acronyms

ACP	Assessment coordinating partner
ASER	Annual Status of Education Report
AP	Assessment partner
DRC	Democratic Republic of the Congo
ECW	Education Cannot Wait
EGRA	Early Grade Reading Assessment
EiEPC	Education in emergencies and protracted crises
EOI	Expression of interest
FER	First Emergency Response
GTP	Global technical partner
HLOM	Holistic Learning Outcome Measurement
ICC	Item characteristic curve
IDP	Internally displaced person
IRT	Item response theory
MYRP	Multi-Year Resilience Programme
SDG	Sustainable Development Goal
SEL	Social-emotional learning
SEACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
UN	United Nations

Glossary of relevant actors

Education Cannot Wait

Education Cannot Wait (ECW) is the global fund for education in emergencies and protracted crises within the United Nations.

Global technical partner

The global technical partner (GTP) provides technical services and oversight to the programme partners on the design and implementation of the Holistic Learning Outcome Measurement (HLOM) Initiative. The GTP facilitates the joint design decision making for the assessment, supports with the selection of the assessment partner, and provides qualitative assurance and assessment related technical support to the assessment partner such as methodological guidance, tool development, data analysis and benchmarking of results. The GTP also supports the programme and sector partners in disseminating the findings, conclusions and recommendations in each country. At the global level, the GTP can lead the development of global public goods for the sector and disseminate learnings from the studies conducted. For ECW's initiative on HLOM, the GTP was Cambridge Education and Oxford MeasurEd.

Multi-Year Resilience Programme grantee

The Multi-Year Resilience Programme (MYRP) grantee, contracted by ECW, is the partner who is responsible for coordinating and managing the funds and implementation of the programme. The grantee is the first point of contact for the GTP and supports with the implementation of the learning assessment. The grantee is also responsible for advocating for the learning assessment project and supports the uptake of findings.

Through its MYRP funding window, ECW invests in countries affected by protracted crises, offering a comprehensive and sustainable approach with multi-annual funding commitments usually lasting three or more years.

Assessment partner

The assessment partner (AP) brings technical measurement expertise in assessment, and provides technical and logistical capacity to implement the assessment. They develop the assessment framework, tools, enumerator training materials, conduct data analysis and write the national report on the findings. They are usually contracted by the grantee.

Assessment coordinating partner

The assessment coordinating partner (ACP) is the partner who maintains the mandate to advocate for learning assessment systems strengthening within crisis contexts broadly, and the MYRP/programme context specifically within a country. The ACP looks for opportunities to advocate for the learning assessment project and supports the uptake of findings. The ACP will usually be the MYRP grantee.

Contents

Tables

Figures

NAVIGATING THE HANDBOOK

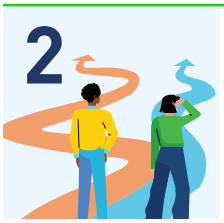
Introduction, purpose and use of this Handbook

Overview of the background and purpose of the Handbook, how to use it, and section on addressing risks and systemic barriers to developing HLOM in crisis settings

Starting the design process

Activities undertaken to reach this phase of design, culminating with the global technical partner and assessment partners leading on design using this Handbook

- **1.1** Identify technical support needed
- **2.2** Kick-off meeting and assessment partner selection
- **2.3** Theory of change workshop
- **2.4** Development of the design outline



Design decisions

Key decisions that partners will need to make before beginning the process of developing an assessment

- **3.1** Policy goal and research questions
- **3.2** Resources available against the ambition
- **3.3** Efficiency and cost drivers
- **3.4** The assessment system to be strengthened
- **3.5** School, learning centre or community-based?
- **3.6** Target grades or ages
- **3.7** Domains
- **3.8** Timing
- **3.9** Contextual characteristics of interest
- **3.10** Evidence use and uptake

Design and execution

Key steps that are common to the development of all assessments

- **4.1** Assessment framework
- **4.2** Language
- **4.3** Items
- **4.4** Test development
- **4.5** Background information
- **4.6** Coverage
- **4.7** Standardization and field operations
- **4.8** Data management
- **4.9** Piloting
- **4.10** Psychometric analysis
- **4.11** Benchmarking
- **4.12** Secondary analysis
- **4.13** Reporting
- **4.14** Dissemination

INTRODUCTION, PURPOSE AND USE OF THIS HANDBOOK

Background

Nearly a quarter of a billion children and adolescents (ages 3-18) are affected by armed conflict, forced displacement, climate-induced disasters and other crises, and in need of education support.[1] There is limited evidence on the learning outcomes of these children. Existing evidence is largely project-based and its overall quality is moderate at best.[2]

School-based assessments administered by – whether through national exams or sample-based assessments at the national, regional or global levels – do not capture data for out-of-school children and are often not representatively implemented in education in emergencies and protracted crises (EiEPC) settings. Organizations working to provide education in crisis contexts are often not sufficiently incentivized and organized to systematically collect data on learning outcomes over time to monitor trends.

Despite the needs of children affected by crisis, there are also fewer initiatives measuring social-emotional learning (SEL) in EiEPC settings compared to academic learning outcomes. As a result, there is limited in-country capacity to identify key SEL domains, define contextually relevant SEL constructs and carry out meaningful measurements.

The purpose of this Handbook is to provide a structure for the design and execution of a tailored holistic learning outcome measurement. The Handbook takes a technical, capacity-development and research uptake approach that aims to strengthen learning assessment systems in crisis settings.

This Handbook was developed as part of the Education Cannot Wait (ECW) Holistic Learning Outcomes Measurement (HLOM) programme. The programme is implemented by a global technical partner (Cambridge Education and Oxford MeasurEd), six ECW Multi-Year Resilience Programme (MYRP) grantees – in Uganda, Iraq, Democratic Republic of the Congo, Burkina Faso, Ethiopia and Bangladesh – and six National Assessment Partners. The Handbook aims to increase the availability of quality holistic learning outcomes data, improve the design and execution of holistic learning outcome measurement in EiEPC contexts, and support system strengthening.

The Handbook was initially developed in 2021 to guide programme partners through the design and implementation of learning assessments across six EiEPC contexts. It was shared with country-level stakeholders before beginning country design processes and was adapted based on experiences across contexts.

This 2025 edition incorporates updates based on pilot implementations, feedback and lessons learned throughout the process.

1 https://www.educationcannotwait.org/resource-library/new-global-estimates-study
2 Outhred, R and Turner, F. UNHCR, Oxford MeasurEd and Mott Macdonald. Education Evidence Brief 2022-2023. https://www.unhcr.org/africa/sites/afr/files/legacy-pdf/634fc3b74.pdf

Use of this Handbook

This Handbook is intended to guide EiEPC programmes and organizations – including ECW grantees – who conduct an HLOM. It outlines the approach used in the ECW-funded HLOM programme and serves as a guide for other organizations conducting an HLOM for the first time or as part of a repeated effort.

The Handbook provides a structured framework for the assessment development process and outlines the steps each partner should follow during the design phase. It focuses the technical aspects of assessment development.

To replicate this process, it is recommended that programmes identify a partner to provide technical oversight, leadership and quality assurance – similar to the role played by the global technical partner (GTP) in the HLOM programme.

Given the wide range of roles and responsibilities across different EiEPC contexts, this Handbook uses specific terminology to describe the various actors involved in the process. While alternative terms may be used, it is important that all stakeholders have a collective understanding of each actor's role and responsibilities.

In the ECW-funded HLOM programme, partners varied across contexts (with the exception of the GTP). For example, assessment partners included government bodies, national NGOs, and national or international consultancy firms.

TABLE 1. TERMINOLOGY FOR DIFFERENT ACTORS

Term	Description
Assessment coordinating partner	The assessment coordinating partner (ACP) is the partner who maintains the mandate to advocate for learning assessment systems strengthening within EiEPC contexts broadly, and the MYRP/programme context specifically, within a country. The ACP looks for opportunities to promote the learning assessment project and supports the uptake of findings. The ACP will usually be the programme grantee.
Assessment partner	The assessment partner (AP) brings technical expertise in assessment and provides technical and logistical capacity to implement the assessment. The AP and ACP may be the same or different.
Strategic partners	Strategic partners include all the partners to be engaged and consulted with throughout the assessment and will usually include government. Strategic partners may or may not be involved in advocating for learning assessment systems strengthening in EIEPC contexts but must be strategically engaged by the ACP over the long term on alignment (where possible) of the system and uptake of evidence emerging from the assessment.
Global technical partner	The global technical partner (GTP) for ECW was Cambridge Education with Oxford MeasurEd. However, for grantees wishing to replicate the process undertaken in the HLOM programme, a consultant or organization needs to be contracted to fill this role. The GTP provides support across a variety of activities, including policy guidance, coordination and partnerships strengthening, technical training, methodological guidance notes, costings, item development, piloting tools, data analysis and advocacy, communication and dissemination. The GTP's role includes supporting the selection of the right AP, if one is required.

Common challenges and how to navigate them

The development and implementation of HLOM in crisis settings are fraught with challenges, many that are foreseen and others that are not. In implementing HLOM programmes across six countries, these challenges ranged from the absence of a standardized framework to the variability in partner capacities and the lack of reliable population data. However, by addressing these challenges through tailored approaches, capacity building and technical solutions to data limitations, it is possible to conduct high-quality learning assessments that provide valuable insights into the educational needs and outcomes of children in emergency settings. This, in turn, can inform more effective and targeted interventions – ultimately improving the educational experiences and prospects of these vulnerable populations.

Four major systemic barriers to developing holistic learning outcome measurements in EiEPC settings were encountered during ECW's HLOM initiative:

1 Absence of a blueprint or framework on how to do an HLOM in EiEPC settings

Very few initiatives measure holistic learning outcomes in EiEPC and, where they do, the initiatives are highly projectized, the geographical scope is narrow and the quality is moderate at best. This lack of an established framework for HLOM in EiEPC settings required the creation of new approaches and methodologies.

The diversity of conflict type, education provision mechanisms, education providers, curriculum exposure, linguistic diversity and security situations meant a specific blueprint for how to do assessment in EiEPC was neither possible nor desirable. Rather, a framework on how to make key decisions to design a holistic learning assessment within each context was needed.

To fill this gap, the Handbook was developed and piloted to guide in-country partners through a decision-making pathway for designing context-appropriate learning assessments.

2 Often unrealistic planning and budgeting for learning assessments

In the context of education in emergency settings, budgeting for children's learning assessments presents several significant challenges. Learning assessments are often regarded as an afterthought, seen as an add-on rather than an integral part of regular programming. Monitoring tends to focus on inputs rather than outputs, and the monitoring of outcomes is rare. Consequently, this leads to insufficient financial planning and allocation for assessing children's learning levels and using that data to inform programming.

Grantees often underestimate the workload required, and so often do not allocate sufficient budget or staff time within their monitoring budgets and workplans.

Moreover, there can be limited understanding among grantees and partners on how to effectively build a budget for quantitative assessments. This lack of expertise can result in poorly constructed budgets that fail to account for all necessary components of an assessment (such as data collection, analysis and

reporting), which can, in turn, result in poor quality delivery from commissioned assessment partners. Annex E provides a budget template for the APs costs.

Additionally, the capacity to quality-assure and critically evaluate the technical proposals and budgets from APs is lacking. This gap in capacity means that grantees may accept technical proposals that will not produce high-quality data and/or financial proposals that are either inflated or insufficient – leading to either wasted resources or underfunded assessments that cannot meet their objectives.

3 Technical challenges not unique to the EiEPC setting

Discussions about conducting holistic learning assessments often start with an incorrect emphasis on selecting tools rather than defining the purpose of the assessment. This approach prioritizes the 'how' over the 'what' and 'why'.

The primary concern should be to first clearly define the purpose and context of the assessment. This involves deciding: (1) what information is needed, by whom, for what purpose and when; (2) what research questions needs to be answered; (3) what resources are available to support the exercise; (4) if the assessment should be school-based or community-based; (5) which ages or grades should be assessed; and, only then, (6) what learning domains will be targeted and therefore, which tools might be appropriate.

4 Technical challenges unique to the EiEPC setting

Conducting learning assessments in EiEPC settings presents several unique technical challenges.

- **Identifying a sample frame:** Identifying a sample frame is particularly difficult due to the high mobility of target populations, the lack of comprehensive population data, and the complexity of determining the unit of analysis in settings with diverse formal and informal education provisions. In many conflict-affected countries, census data on refugees, internally displaced persons (IDPs), returnees and host communities are often unavailable. Population information is typically gathered at the project level or through the education cluster, which also depends on data from individual projects.
- **Quality and type of enumerators:** Recruiting quality enumerators and supervisors is challenging due to security concerns, accessibility to learning spaces, diverse language requirements for gathering data from displaced populations, and limited budget allocations for data collection in crisis settings. Quality data collection in emergency settings takes longer and therefore relatively costs more compared to non-crisis settings where learning spaces are more accessible. In Iraq, due to budget limitations, MYRP community members with experience in data collection were used as enumerators and trained online by the national AP; in Burkina Faso, government officials collected the data; and in Bangladesh, the AP had to recruit individuals from Cox's Bazar that spoke Burmese but had limited experience as enumerators. To mitigate the risks associated with inexperienced enumerators, it is crucial to provide comprehensive training and appoint a field manager to oversee the data collection process. In addition, data should be monitored "live" during fieldwork so that any issues with data quality can be addressed before enumerators return. This approach ensures that enumerators are well-prepared, and that the data collection is conducted efficiently and accurately under experienced supervision.

TABLE 2. COMMON CHALLENGES IN CONDUCTING HLOM PROJECTS AND MITIGATION STRATEGIES

Common challenges in conducting HLOM projects	Mitigation
The absence of a framework on how to conduct an HLOM in EiEPC settings	Use this Handbook to provide the framework for designing the assessment
Unrealistic planning and budgeting for learning assessments	Ensure you have accounted for the full costs of conducting an HLOM project. Costs include: • Staff time to manage and coordinate partners and workplan the process • Contracting for national experts (such as leading education academics) to participate in workshops, and review and validate findings • Contracting a GTP • Contracting an AP Appoint one person within the ACP to be the single line of communication with all partners, holding responsibility and designated time within the role for: • Leading procurement • Coordinating partners • Creating and managing the workplan • Providing all the information needed by APs
Technical challenges not unique to the EiEPC setting:	
• **Not leading with the purpose and context of the assessment**	This Handbook provides sequencing for making design decisions. The GTP can support partners to follow this sequencing and avoid leading with "which tool should we use"
Technical challenges unique to the EiEPC setting:	
• **Challenges in identifying a sample frame**	Begin gathering data immediately from implementing partners, using the template provided in Annex A (suggest cutting and pasting it into an excel file)
• **Quality and type of enumerators**	Consider how enumerators will be selected, recruited, trained and managed. Several options are available, including: • Include enumerator hiring, managing and training in the AP terms of reference and contract • Use local community members, trained and managed by the AP

1
STARTING THE DESIGN PROCESS

1.1 IDENTIFY TECHNICAL SUPPORT NEEDED

A first step in the design process is to identify and appoint an internal or external organization to fill the role of GTP. The terminology 'global' in GTP does not imply that the expert organization needs to reside in any particular location, but rather that they have a set of knowledge and skills that is globally recognised in the learning measurement space. The GTP can be a national or international organization. It can also be a consultancy firm, university, headquarters or regional office from the United Nations or an NGO that is capable of providing the technical support and advice to their country programmes.

Table 3 outlines the basic tasks that should be undertaken by a GTP, if the national AP has a lot of experience conducting learning assessments. These are tasks that UN or NGO country offices often do not have (or need to have) in-house.

TABLE 3. BASIC RESPONSIBILITIES OF THE GTP

Tasks completed by the GTP	☑ Manage the inception phase and contracting of technical support
	☑ Identify systems change indicators and tracking changes
	☑ Undertake sampling to inform terms of reference and procurement process for national AP
	☑ Support in drafting the terms of reference for the national AP
	☑ Review technical proposals, attend any interviews, and provide scoring and feedback
	☑ Prepare for theory of change and design decisions workshop
	☑ Facilitate theory of change and design decisions workshop
	☑ Undertake capacity assessment of national assessment partner
	☑ Review assessment tools for face validity
	☑ Review contextual data tools

continued: **Tasks completed by the GTP**	☑ Undertake psychometric analysis of pilot data and provide feedback
	☑ Undertake data checks after fieldwork
	☑ Conduct psychometric analysis on full dataset
	☑ Conduct benchmarking in collaboration with stakeholders
	☑ Review final report and provide feedback
	☑ Support with dissemination

Table 4 lists additional activities that are required if the AP needs additional support.

TABLE 4. ADDITIONAL SUPPORT TASKS FOR THE GTP

Additional tasks completed by the GTP when the national partner needs additional support	☑ Update additional rounds of sampling as population data becomes more complete
	☑ Adapt or create assessment tools
	☑ Adapt or create contextual data tools
	☑ Fieldwork monitoring
	☑ Analysis
	☑ Report writing

To select and contract a GTP, clear selection criteria need to be applied to ensure the GTP has the right set of skills and experience to be able to provide technical expertise, and that the investments in the assessment process will result in high-quality outputs. Table 5 shows the selection criteria that the ECW-funded HLOM programme applied.

TABLE 5. GTP SELECTION CRITERIA

Suggested GTP selection criteria		Means of verification	Details
1	Willingness and interest to undertake the work	Consultations with in-country grantees	Yes/No
2	Prior experience in designing learning assessments in EiEPC settings, conducting capacity assessments of national assessment partners, conducting psychometric analysis, and analysing and reporting on learning outcomes	Expression of interest (EOI) from individuals/ organization	Description
3	Prior experience in developing sampling strategies in emergency contexts	Specific project references in EOI from organization	Description
4	Demonstrated in-house capacity to oversee the technical work of data collection agencies	EOI from organization, CVs of senior staff and enumerator mobilization processes (recruitment), and an interview	Description
5	Experience working with and through governments, especially in advocating for education in emergencies issues and presenting complex data to a lay audience	EOI from organization	Description
6	Experience of evidence uptake with key stakeholders	EOI from organization, examples of reports and uptake of results	Description
7	Disclosure of any potential conflict of interest	Signed declaration from organization	Yes/No
8	Established track record of working in and making meaningful contributions within the wider education sector	References	Description

Based on experiences in designing and implementing holistic learning outcome measurements in EiEPC settings through this programme, there are four important steps to begin the design process once the GTP is in place. The following section provides more information about each of these steps.

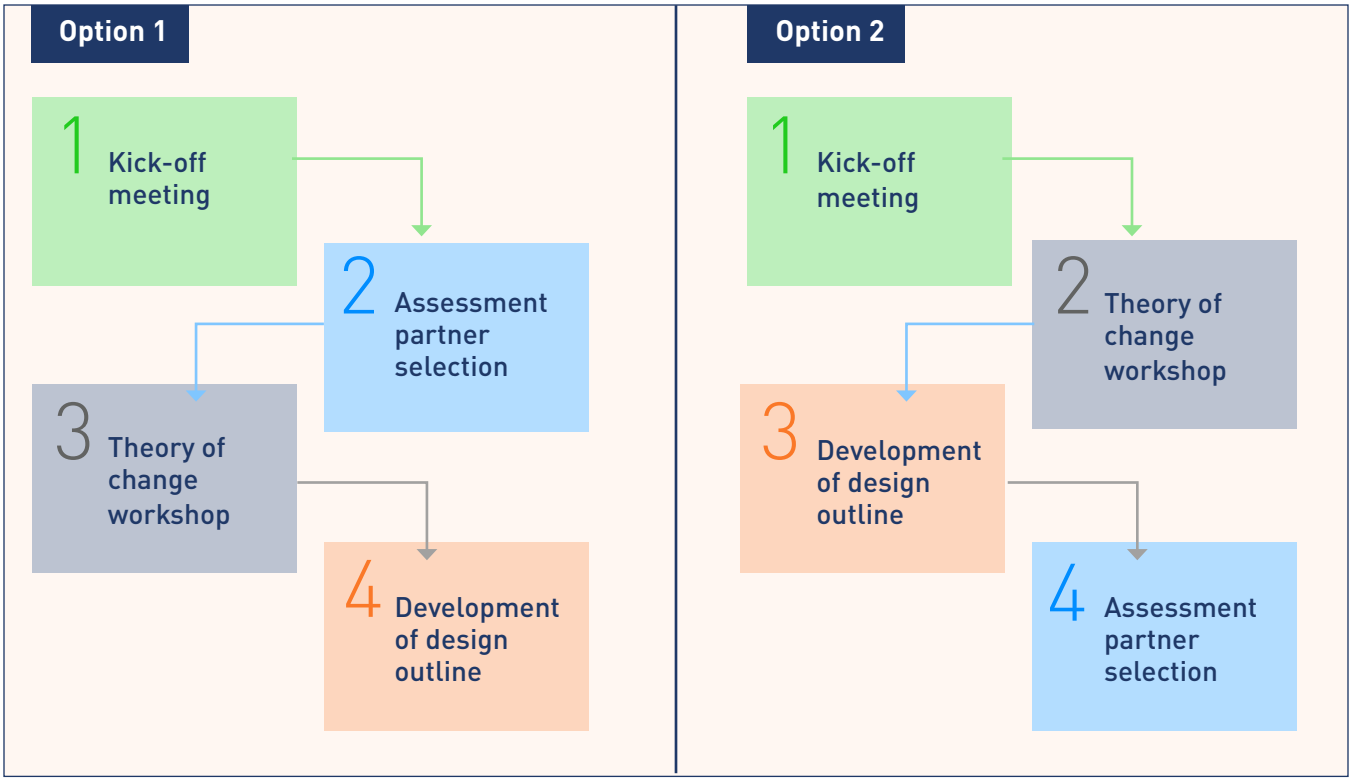
1.2 KICK-OFF MEETING AND ASSESSMENT PARTNER SELECTION

The design phase should begin with a design kick-off meeting between the ACP and the GTP to explore expectations and understand existing capacity need.

Based on the internal capacity of the organization and programme team, a national AP is to be identified, selected and contracted. The length of the procurement process will determine if it is feasible to wait for the AP to be onboarded before the GTP facilitates the theory of change workshop and a design decisions workshop. The advantage of undertaking the design decisions workshop first is that these decisions may inform the terms of reference for the AP.

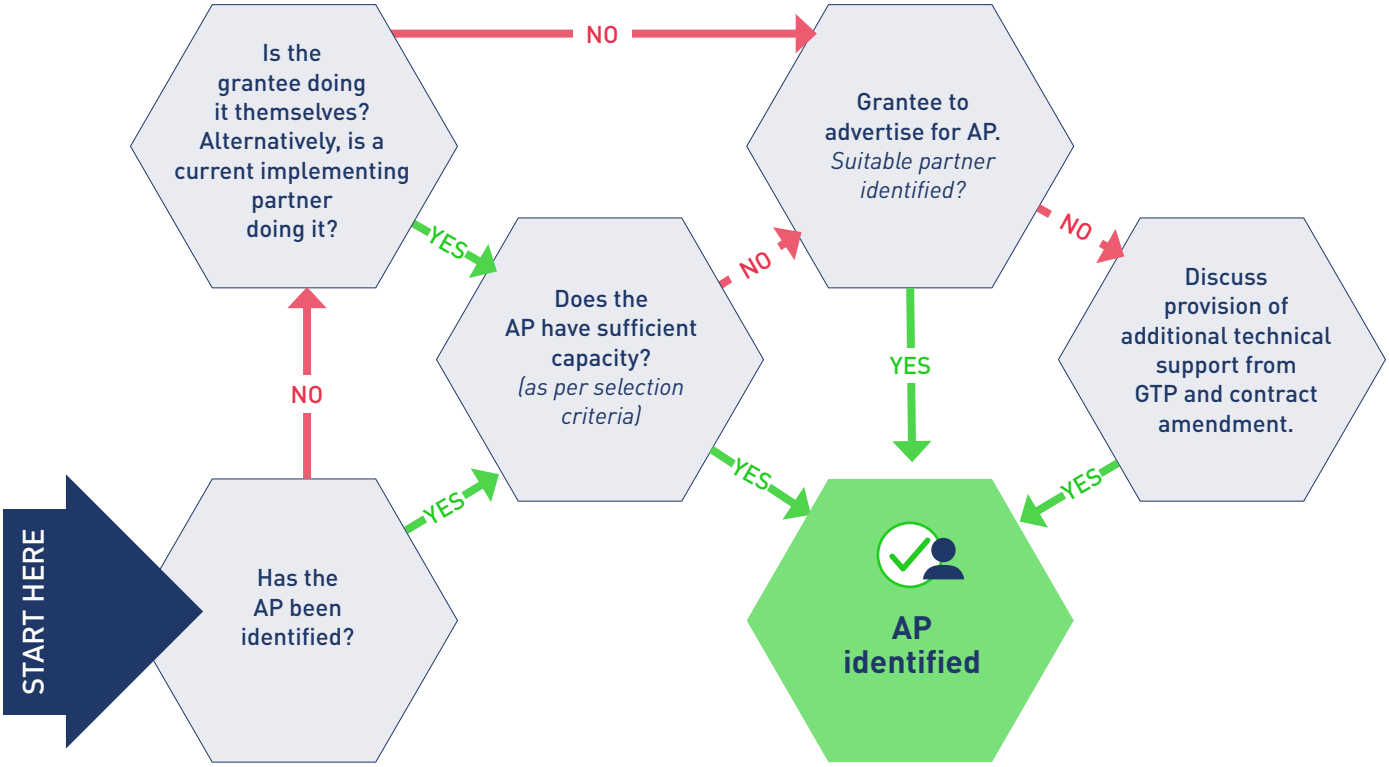
Figure 1 and 2 provide the two options regarding the sequencing of steps:

FIGURE 1. FIRST STEPS IN THE DESIGN PROCESS



To ensure an AP is in place at the appropriate time, the process to identify an AP can begin at the beginning of the design phase. The following decision tree can be used to identify an AP.

FIGURE 2. ASSESSMENT PARTNER IDENTIFICATION DECISION TREE



To acquire the services of a high-quality AP efficiently through a competitive process, consider the following learnings and recommendations:

- Develop clear terms of references that include the design decisions tracker, sampling strategy and budget template (provided in Annex E).
- Identify the best platforms to advertise the request for proposals to reach a wide national and international audience.
- Set a deadline for clarification questions and respond in a timely manner.
- Identify qualified reviewers for the technical proposals – assessment specialists who can critically evaluate the methodologies proposed and related budgets.
- Set realistic timelines for the entire procurement process and assessment start dates.

The GTP supports and provides oversight to this process.

Table 6 outlines the suggested selection criteria and means of verification used in the ECW funded HLOM programme.

TABLE 6. ASSESSMENT PARTNER SELECTION CRITERIA

	Suggested in-country AP selection criteria	Means of verification	Details
1	An organization that manages or implements the programme	Approved programme	Yes/No
2	Willingness and interest to undertake the work	Consultations with in-country grantees	Yes/No
3	Prior experience in conducting learning assessments	EOI from organization	Description
4	Prior experience in conducting representative sample surveys in emergency context	Specific project references in EOI from organization	Description
5	Demonstrated in house capacity of the team to lead a survey and ability to mobilise necessary resources	EOI from organization, CVs of senior staff and enumerator mobilization processes (recruitment) and interview	Description
6	Experience of working with and through governments, especially in advocating for education in emergencies issues and in capacity development	EOI from organization	Description
7	Experience of evidence uptake with key stakeholders	EOI from organization and examples of reports and uptake of results	Description
8	Disclosure of any potential conflict of interest	Signed declaration from organization	Yes/No
9	Established track record of working in and making meaningful contributions within the wider education sector	References	Description

2.3 THEORY OF CHANGE WORKSHOP

The GTP will next facilitate a theory of change workshop[3] with the ACP and strategic sectoral partners to adapt the programme's global theory of change to the country setting and align with the in-country programme. The theory of change workshop explores country-level objectives of the assessment, introduces the theory of change methodology, jointly considers the most important contextual factors that can affect implementation, confirms that the right activities are being undertaken to support the overall goal, and examines challenges and refines causal pathways leading to outcomes.

The potential barriers and enables will also be explored to 'stress test' the theory of change. The workshop may take place remotely or in-person – however, experiences suggest that face-to-face theory of change and design decisions workshops create much greater traction and buy-in with partners, particularly government partners.

The theory of change could also include system strengthening work. The GTP will work with country stakeholders to identify the perceptions, incentives and barriers to change for each of the stakeholders involved in the education system. The theory of change will be sense-checked to ensure the envisaged change in the system is realistic. The SABER framework outlined in Section 3.3 can provide the conceptual basis for systems change discussions as the GTP works with the ACP to identify desired changes and track progress. Where barriers to system-level change are identified in the theory of change workshop, these will be documented and a wide range of stakeholders will be engaged to discuss how these can be addressed.

The workshop should focus specifically on the outputs and outcomes link in the results chain to understand:

- Overall goal of the assessment and systems strengthening work
- Steps required to develop and implement holistic learning outcome measurement and systems strengthening
- Barriers towards the use of learning outcome data when it is in place within the EiEPC setting
- What needs to be addressed to ensure learning outcome data is both available and used to inform decision-making within the larger theory of change for the setting

3 Materials for this have been provided as part of this handbook package.

2.4 DEVELOPMENT OF THE DESIGN OUTLINE

The main counterpart for the GTP is the ACP (most likely the programme grantee or organization that implements the education programme). Design decisions regarding the assessment and measurement system strengthening should be ultimately made by the ACP, after a consultation process with strategic partners and with guidance from the GTP. Ultimately, the design may need to be signed off by the programme's Steering Committee or Education Cluster. The Steering Committee will be responsible for ensuring programme activities align with the appropriate frameworks in each context (for example, Education Cluster plans, the Humanitarian Response Plan, national priorities and the national education sector plan).

The GTP should engage with the ACP regarding the consultation process for the design outline development. The ACP should be encouraged to lead the consultation process with strategic partners using this Handbook, and with support and backstopping from the GTP. The GTP should participate in consultations and provide expert views on design decisions, including potential trade-offs, likely costs and other technical know-how.

The design outline should include decisions on the aspects of the assessment, outlined in the following phase.

2 DESIGN DECISIONS

At the start of the learning assessment design process, it is important to establish the parameters for the assessment. These are set out in Figure 3.

FIGURE 3. SUMMARY OF KEY DESIGN DECISIONS

A table is included in Annex C to document these decisions. The table should be completed in consultation with the main users of the information that should be provided by the learning assessment. They will need some time and guidance to consider carefully what evidence they need. This Handbook provides information on how to approach some of these decisions and should be consulted regularly. An important principle throughout is to probe until answers are as specific as possible. Without this, the designers of the learning assessments will need to fill in gaps or make decisions without clear guidance, and the result may fail to achieve what the intended users had hoped for.

The process should start by identifying the primary users of the learning assessment as well as the owner (usually the responsible government body or organizations responsible for providing education services). This should include workshops to allow for different thoughts and ideas to interact, but may also include interviews with key decision-makers to ensure their needs are well understood.

2.1 POLICY GOAL AND RESEARCH QUESTIONS

KEY POINTS:

- For an assessment to be 'fit for purpose' or 'valid', the purpose needs to be agreed on and clearly articulated
- The policy goal should not be broad or vague (e.g. 'provide information about…') but, rather, should be as specific as possible and meet the purpose set (e.g. 'ensure children are placed in appropriate grades based on their current learning levels')
- It may be possible that a small number of well-defined policy goals can be articulated, however, no single assessment can respond to all purposes well
- From policy goals, specific research questions that will inform the policy goal need to be agreed on – this will inform decisions regarding which data needs to be collected
- We use the case of the ASER Plus initiative in Cox's Bazar, Bangladesh (see page 27) to explore how to explicate the policy goal during the design phase

For a learning outcome assessment to be 'fit for purpose' or 'valid', the purpose needs to be agreed on and clearly articulated. The decisions that make the assessment suitable for its intended purpose may make it unsuitable for other uses or inferences. Additionally, the decisions that will be made during the design of the test instrument should be tailored towards this purpose, so it is important that this is considered carefully and collectively.

It is also important to consider if learning outcome information is to be gathered at the individual learner, group or population level.

Note on Equity: Learning assessments for equity assess if and to what extent any inequities exist between different learner groups (e.g. gender, displacement status, nationality, ethnicity, disability, home language, etc.). If inequities in learning outcomes do exist, assessment findings can provide data that can help you identify barriers faced by different learner groups.[4]

After identifying the purpose, a specific policy goal and research questions are to be identified. These should not be broad or vague but, rather, should be as specific as possible. There can be a tendency towards vagueness when building consensus because it can provide a form of words that all sides can agree on. However, this should be avoided because, in this case, it will only push disagreements to a later point when nothing can be done to address the issues. Instead, consensus should be built by:

1. Identifying a long list of possible purposes
2. Prioritising the purposes that will have greatest impact on student learning and well-being
3. Identifying decisions that are compatible because they require the same information (i.e. the same formative/summative category, grades/ages, domains, timing) and listing these purposes explicitly, when possible.

In short, be specific and prioritise rather than attempting to capture multiple purposes within one vaguely worded one.

It may be possible that a small number of well-defined policy goals can be articulated, however, no assessment can meet all policy goals.

The purpose might be to support a process, action or decision. For example, it may be to assist funding decisions, identify areas of programme successes and failures (evaluation judgements), to identify specific learning needs and/or variation in outcomes between groups and sub-groups.

From policy goals, specific research questions that will inform the policy goal must be agreed on. This should inform decisions regarding which data needs to be collected. Table 7 outlines questions for which the answers will help identify the policy goal.

TABLE 7. POLICY GOAL QUESTIONS TO ASK AND BE ANSWERED

Questions	Answers
1. What decision(s) will be made using the assessment results data?	
2. Who will make the decisions?	
3. When will decisions be made?	
4. What information is required that is not available currently?	
5. How does this information support the decisions or uses?	
6. About which people or groups do you need the learning assessment information?	

4 https://resourcecentre.savethechildren.net/pdf/SC-Measuring-Learning-Outcomes-Guidance_FINAL.pdf/

An example of how these questions can be used to shape the design of an assessment is shown in Figure 4. This is based on the case study of ASER Cox's Bazar, detailed later in this section. The research questions the ASER study was able to answer are then listed with links to the case study policy goals (Table 9).

FIGURE 4. EXAMPLE OF POLICY GOAL FORMULATION PROCESS

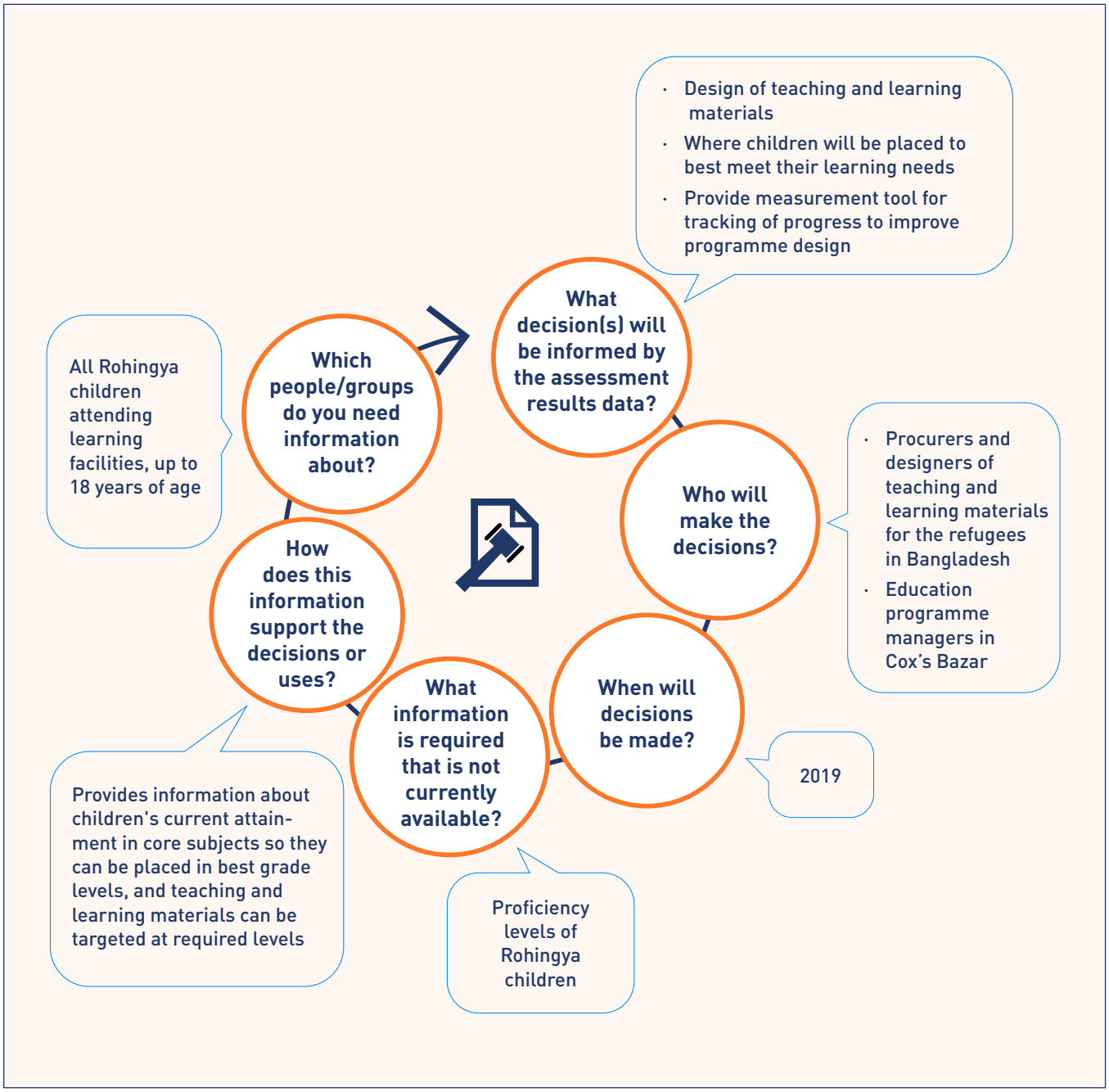


Table 8 shows some examples of assessments and their purposes to illustrate what they might look like.

TABLE 8. EXAMPLES OF ASSESSMENTS AND THEIR PURPOSES

Assessment type	Purpose
Uwezo	The purpose of the Uwezo learning assessment is to: • "Contribute to the improvement of the quality of education. Annual household surveys are implemented to assess the basic literacy and numeracy competencies of school-age children across Kenya, Tanzania and Uganda. Uwezo believes that this information will raise public awareness about education levels and will trigger actions aiming to improve them."[5]
Southern and Eastern Africa Consortium for Monitoring Educational Quality (SEACMEQ)	The purpose for SEACMEQ studies are to: • "Provide educational officials and researchers with training in the technical skills necessary to monitor, evaluate, and compare the general conditions of schooling and the quality of basic education • Generate reliable information that can be used by decision-makers to formulate effective plans aimed at improving the quality of education • Widely disseminate and promote the use of SEACMEQ research findings as a foundation for evidence-based policy and practice."[6]
Early Grade Reading Assessment (EGRA)	EGRA can serve several purposes, including: • "Serve as a baseline of early reading acquisition • Guide the content that is included in an instructional programme • Be used to evaluate programmes."[7] Due to its informed theoretical framework and consistent procedures, EGRA yields valid and reliable data for all the purposes outlined above. Additionally, EGRA offers a common language for discussing children's literacy abilities. For example, results from the passage reading subtask allow us to gauge how well children can handle grade-level text.

5 Australian Council for Educational Research (2014) *Uwezo: Monitoring children's competencies in East Africa*, Assessment GEMS Series, No. 7. Available at: https://www.acer.org/files/AssessGEMs_Uwezo.pdf

6 Australian Council for Education Research (2015) *The Southern and Eastern Africa Consortium for Monitoring Educational Quality*, Assessment GEMS Series No. 8. Available at: https://www.acer.org/files/AssessGEMs_SACMEQ.pdf

7 Dubeck, M. and Gove, A. (2015) The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations *International Journal of Educational Development* Vol 40, pp. 315-322

CASE STUDY
The ASER Plus Initiative in Cox's Bazar, Bangladesh

ASER Centre, the survey and research arm of Pratham Education Foundation in India, has developed a basic literacy and numeracy tool (ASER tool) used in the country for the annual, citizen-led national household survey of basic skills since 2005.

In 2018, funded through the ECW MYRP I, UNICEF – as the education cluster coordinator – contracted the INGO Room to Read to adapt the ASER tool to the context of Rohingya refugees in Cox's Bazar.

Objective:

The main objective was to assess the learning levels of enrolled children based on LCFA Levels I–IV so to:

- Place children with LCFA proficiency levels (diagnostic)
- Set the baseline of enrolled Rohingya children on literacy and numeracy levels
- Obtain a sense of direction regarding the preparations of teaching and learning materials
- Advise education sector partners on teaching and learning interventions that respond effectively to the learners competency level

Tool/Method:

ASER Plus, utilizing the contextualized ASER tool, was conducted in learning facilities as opposed to households. The population-based survey covered the children attending the learning facilities, up to 18 years of age. The survey assessed children's knowledge of Burmese and English language, and mathematics. Data collection started on 3 December 2018 and was completed on 13 December 2018. All tools, data collection sheets and survey guides are available in the ASER Plus report by Room to Read.

Since it was a population survey for diagnostic purposes, a huge effort was made to assess all the children in the learning facilities. A total of 180,470 (52% boys and 48% girls) – enrolled in 2,499 child learning facilities located within the camps, supported by 13 national and international NGOs – were tested. 67 master trainers trained 5,000 teachers to carry out the ASER Plus survey.

TABLE 9. EXAMPLE OF RESEARCH QUESTIONS THAT INFORM THE POLICY GOAL (Cox's Bazar, Bangladesh)

Specific research questions	How these questions will inform the policy goal
What are the literacy and numeracy levels of individual students in Cox's Bazar, against Learning Competency Framework and Approach proficiency bands?	Place children in appropriate learning environments
What are the baseline literacy and numeracy levels of refugee children in Cox's Bazar?	Support understanding of the baseline learning levels of refugee children to track progress towards Sustainable Development Goal 4
What specific areas of learning do refugee children in Cox's Bazar find difficult?	Inform education sector partnerships for any future programme interventions

2.2 RESOURCES AVAILABLE AGAINST THE AMBITION

KEY POINTS:

- The policy goal to be addressed and research questions to be answered need to be realistically delivered within the level of resources available
- Where there is potential for additional funding that has not yet been secured, design the programme around what is available and plan for additional 'add on' options that new funding sources might be able to support

As early as possible in the process, the resources available to design and implement the assessment and associated capacity-building activities need to be mobilised to bear on the design. There have been many assessment activities that have failed to deliver on any policy goals or research questions because the scale and ambition of the design was not aligned to the available resources.

In some situations, in addition to core funding through the programme, there are potential additional funds that have not yet been secured. In this situation, the design should include both a core design based on available funds, and additional 'add on' options. For example, core funding might support an assessment of literacy, numeracy and social-emotional skills with a sample size that can provide accurate estimates of the learning levels of the overall population of students in a particular grade. An 'option' that can be pursued with additional funding could increase the sample size and allow for precise sub-group estimates of learning levels (allowing comparisons of the learning

levels of students across different camps or districts, for example).

Where the level of ambition is greater than the available resources, policy goals and research questions will need to be revisited prior to moving forward with design work.

2.2.1 Efficiency and cost drivers

Developing assessments that perform their intended purpose efficiently requires an understanding of the drivers of quality or fitness for purpose, and the drivers of costs. This enables designers to make the best trade-offs to minimise costs, while ensuring that the instrument provides sufficient validity.

Efficiency is also achieved by identifying problems early. This can require additional resources early on but saves significant wasted resources later in the process. For example, developing, adapting and piloting excess items requires greater resources early on, but it avoids further rounds of re-writing and piloting. The expended resources are significantly smaller than those required to do additional rounds of development and piloting.

The greatest cost associated with implementing learning assessments is the fieldwork. Large teams of test administrators are required, together with management, transport and accommodation costs. Therefore, it is important to be confident that the instrument will work as required before starting fieldwork (this is the purpose of piloting). It is also important to make instruments and fieldwork models as lean as possible. Collecting surplus information that is not required for the stated purpose wastes time and resources. There should be as few instruments as possible, and the instruments should be as short as possible.

The choice between administering tests one-on-one or to groups largely depends on the nature and format of the test, and the characteristics of the students. However, there are cost implications. Administering to students one-on-one lends itself to the use of technological tools that reduce data capture time later in the process. It is also often a necessity to administer assessments one-on-one when the students being assessed are not yet literate. Administering to groups dramatically reduces the amount of administrator-time per student. The trade-off is that this usually requires the use of paper-based assessments, and the data still needs to be entered into electronic form from the responses written on paper.

Scoring is also required for the full sample of cohort of students who are assessed. This can require significant resources. At the costliest end of the spectrum, marking of long-form answered questions can take a long time and require checks for consistency between markers and over time. At the least costly end, multiple choice can be marked by computers, which particularly saves money where there are large numbers of students.

2.3 THE ASSESSMENT SYSTEM TO BE STRENGTHENED

KEY POINTS:

- Identify assessment system to be strengthened
- Describe any system that currently assesses learning – or could do so in the EiEPC context – as well as the key features of the intended strengthened system
- Identify financial, political and technical challenges so they can be addressed
- The GTP has already undertaken part of this exercise as a part of their country rapid analysis and are able to discuss their findings and recommendations

The assessment system to be strengthened needs to be defined. It may be the national assessment system, an assessment system particular to the emergency response (the 'micro education system') or both. Usually, it should be the system that currently assesses learning, or has the potential to assess learning within emergency and protracted crisis settings. In some countries, support may be provided to the government to expand their learning assessment system to crisis contexts. In other countries, support might be provided to an international NGO, civil society or the United Nations to improve the assessment of learning – with the ultimate goal of integrating learning assessment into the national system. In other countries, integration may not be financially, politically or technically possible, and support should remain with non-government partners.[8]

Being cognisant of how the system will function at the conclusion of the programme is the key to sustainability.

8 For example, the Rohingya population around Cox's Bazar in Bangladesh.

CASE STUDY
Considering the system to be strengthened

In order to assess the bottlenecks to conduct and learn from EiEPC-adapted holistic learning outcomes, the GTP used the Systems Approach for Better Education Results (SABER) Student Assessment tool as the guiding framework. The SABER framework promotes stronger assessment systems that contribute to improved education quality and learning for all by documenting and analyzing student assessment systems in developed and developing countries around the world.

The SABER Student Assessment states that to be useful, assessments need to be strong in three areas:

1. Enabling context (school environment, uptake of evidence, etc.)

2. System alignment (purpose of the assessments in place, applies to EiEPC content or not, etc.)

3. Assessment quality (including type of assessment, academic and SEL assessments, age groups and ex-disaggregation)

The institutional assessment outlined the situation in each country, both in relation to the system in general, and the system in relation to refugees, IDPs and other conflict-affected populations. This assisted in identifying the potential added value of this initiative in comparison with what is already in place, and the extent to which bottlenecks are technical, political or a combination of both.

The table below outlines the areas the SABER Student Assessment uses to assess systems (see SABER Student Assessment for more information).

Areas used to assess systems	Description of the current system	Description of the current system for conflict-affected populations	Description of the desired system for conflict-affected populations
Enabling context			
Policies			
Leadership and public engagement			
Funding			
Institutional arrangements			
Human resources			

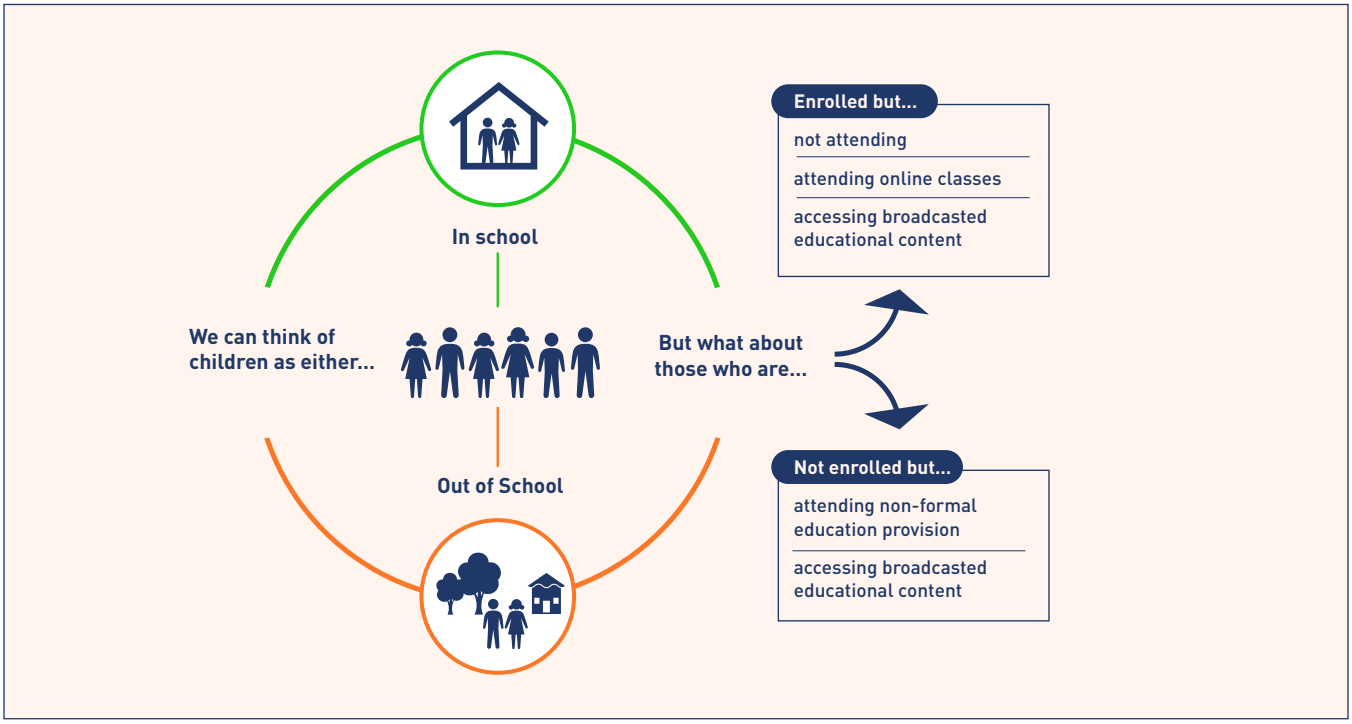
Areas used to assess systems	Description of the current system	Description of the current system for conflict-affected populations	Description of the desired system for conflict-affected populations
System alignment			
Learning/quality goals			
Curriculum			
Pre- and in-service teacher training opportunities			
Assessment quality			
Ensuring quality (design, administration, analysis)			
Ensuring effective uses			

2.4 SCHOOL, LEARNING CENTRE OR COMMUNITY-BASED?

KEY POINTS:

- School or learning centre-based assessments capture information about the learning attainment of students in school or receiving an education intervention
- Community-based assessments capture information about the skills and knowledge of all children – whether they are in formal or non-formal schooling, or not at all
- In emergency and protracted crisis settings, the 'schooling system' is not always formalised or easily defined
- Decisions regarding which population groups should be included and which groups should be excluded will need to be made
- During and following the global COVID-19 pandemic, the lines between who is considered in school and out of school are even more blurred, heightening the need for clear decisions linked to the policy goal

FIGURE 5. BLURRED LINES BETWEEN IN AND OUT-OF-SCHOOL CHILDREN



School or learning centre-based assessments capture information about the learning attainment of students in school or receiving an education intervention. They exclude children who are out of school, thereby leaving a significant information gap. However, they are preferred when the purpose is to understand the effectiveness of the school system or interventions from other education providers. Of course, in emergency settings, the 'schooling system' is not always formalised or easily defined. Schooling can be in formal public or private schools, and it can also be through non-formal providers, such as for accelerated learning or catch-up programmes. Decisions regarding which population groups should be included and which population groups should be excluded will need to be made. Smaller samples of groups of interest might be considered where there is not sufficient funding to survey representative samples for all groups of interest.

Community-based assessments capture information about the skills and knowledge of all children, whether they are in formal or non-formal school, or not. They should be utilized when the purpose is to understand the well-being and development of the whole target population, including those who are out of school. They may be more sensitive to interventions in the community as well as demonstrating learning gains because of improved enrolment.

Community-based assessments may also be preferable where education is provided through a more dispersed delivery model.

During and following the global COVID-19 pandemic, the lines between who is considered in and out of school are even more blurred. For this reason, it is very important during the design phase to have in-depth discussions and to formulate concrete agreements regarding who should be included or excluded from the assessment, and to align these decisions with the policy goal of the assessment.

2.5 TARGET GRADES OR AGES

KEY POINTS:

- Target ages, grades and the programme target group should be selected with reference to the policy goal

Target ages and grades should be selected with reference to the policy goal. For example, if the purpose is to assist or evaluate a programme, the ages or grades should be aligned with the programme being targeted. The characteristics of the target population and of the education system should be considered. Usually, for school-based assessments, target grades should be specified. For community-based assessments, target ages should be specified.

However, within emergency contexts there may be reasons to move away from this tradition and consider different options. There may be a grade level that students need to be prepared for to enter the formal school system and, therefore, the target age may be an age range.

For emergency contexts, this is also the time to determine the programme target group. This may include refugees, IDPs or host communities. It may require defining more precisely the group of refugees or IDPs that will be assessed. These decisions will shape decisions later in the process because different groups have different educational backgrounds, languages, cultures, home situations and vulnerabilities. These need to be accounted for in assessment design to ensure that accurate, meaningful information is elicited. They also need to be considered in the safeguarding processes that are put in place to make sure that all children are protected, and no further harm is inadvertently imposed by the assessments.

2.6 DOMAINS

KEY POINTS:

- Domains are a description of the skills and knowledge that are going to be assessed
- Cognitive and SEL domains should both be considered
- The definition of domains will depend on the context and the age group being assessed

The domains are a description of the skills and knowledge that are going to be assessed. What are the specific learning objectives or outcomes? What skills, knowledge or traits do you want students to demonstrate? This needs to be precisely expressed, preferably with reference to official frameworks.

Often, domains are described in curricula, but in many cases, these do not exist, are not implemented or are out of date. In their absence, more effort and engagement may be required to establish what education providers are trying to achieve or the learning outcomes they should be targeting.

When selecting domains, consider:

- What is being targeted by curricula and/or programmes?
- Which skills are most important to achieve intended impact/results?
- What definition will allow comparisons that you may wish to make?
- Are the skills practical, applied to real life contexts, or attached to curriculum?

Cognitive and SEL domains should both be considered. Technical and vocational skills should also be considered where relevant.

The most common cognitive domain is reading competency, which usually includes decoding and comprehension sub-domains. Literacy is a broader domain than reading, encompassing all forms of obtaining meaning from symbols and writing. Definitions also vary to some degree (TIMSS and PIRLS assessments incorporate an element of social context into their definition).

Assessing against a curriculum:
Determining which curriculum to assess against can be problematic. In some cases, multiple curricula are used across different age groups, or there is no officially recognized curriculum for refugee and IDP children. For example, in Bangladesh, the sector was transitioning from one curriculum to another. In such a scenario, the selection of the curriculum dictates the purpose of the assessment rather than the other way around.

Assessing children on the original curriculum (Learning Competency Framework and Approach) would reveal their current learning levels after being exposed to years of education programming, while assessing them against the new curriculum would establish a baseline for new programming. In Iraq, there was no officially approved curriculum by the Government specifying the skills children should have by a certain grade. In both these scenarios, it meant that children were assessed against international standards for foundational learning. While there are clearly trade-offs for different approaches, in the HLOM programme, we prioritized generating data that remains relevant and provides meaningful insights despite the lack of a standardized curriculum.

TABLE 10. TYPES OF ASSESSMENT TOOLS BY LITERACY DOMAIN AND SUB-DOMAIN[9]

Category	Tools	Reading		Linguistic			Metalinguistic	Writing
		Decoding	Comprehension	Listening	Speaking	Vocabulary	Phonological Awareness	
System monitoring tools								
International assessments	PISA-D, TIMSS, PIRLS, LaNA	–	●	–	–	◔	◔	◔
Regional assessments	LLECE, SACMEQ, PASEC, SEA-PLM, PILNA, UNRWA-MLA	◑	●	◔	◔	●	●	◑
Household	MICS-ECDI, MICS-FLS	●	◑	–	–	–	–	–
Multi-purpose tools								
Foundational skills	MELQO, EGRA, STAR, Literacy Boost	●	◑	◔	◑	◑	◔	◔
Citizen led	ASER, UWEZO, LearnNigeria etc.	●	●	–	–	–	–	–
Designed for EiE contexts	HALDO, IDELA, OLA	●	◑	◔	◔	●	◑	◔

● All tools in category include sub-domain

◑ Half or more of tools in category include sub-domain

◔ One or less than half of tools in category include sub-domain

– NO tools in category include sub-domain

9 Anderson, K., Read, L., Loseda, E., (2020) Academic Learning Measurement and Assessment Tools in Education in Emergencies: Identifying, Analysing, and Mapping Tools to Global Guidance Documents. INEE. Page 57. https://inee.org/system/files/resources/Academic%20Learning%20Measurement%20and%20Assessment%20Tool s%20in%20Education%20in%20Emergencies.pdf

Similarly, numeracy definitions can vary but tend to include number knowledge including basic operations. International assessments include various combinations of other sub-domains, such as math proficiency (e.g. problem solving), measurement, statistics, geometry and algebra.

TABLE 11. TYPES OF ASSESSMENT TOOLS BY NUMERACY DOMAIN

Category	Tools	**Math proficiency** (e.g. problem-solving, reasoning)	**Number knowledge** (Including operations)	**Measurement**	**Statistics and probability**	**Geometry**	**Algebra**
System monitoring tools							
International assessments	PISA-D, TIMSS, PIRLS, LaNA	◐	●	◐	●	●	◐
Regional assessments	LLECE, SEACMEQ, PASEC, SEA-PLM, PILNA, UNRWA-MLA	●	●	●	◐	◐	◐
Household	MICS-ECDI, MICS-FLS	–	●	–	–	–	◔
Multi-purpose Tools							
Foundational skills	MELQO, EGMA, Numeracy Boost	◔	●	◔	–	◔	◔
Citizen led	ASER, UWEZO, LearnNigeria, etc.	◔	●	–	–	–	–
Designed for EiE contexts	HALDO, IDELA	◔	●	◔	–	◔	–

● All tools in category include sub-domain

◐ Half or more of tools in category include sub-domain

◔ One or less than half of tools in category include sub-domain

– NO tools in category include sub-domain

The definition of domains will depend on the context and the age group being assessed.

Domains may be defined to reflect what is taught in schools and learning centres, or they may have more of a focus on practical uses in real-world situations. The age of the target population will also influence the definition as different sub-domains receive greater and lesser focuses at different stages of education (e.g. statistics receives more attention in later grades, while telling time may receive more attention in earlier grades).

The Inter-agency Network for Education in Emergencies defines SEL as "the process of acquiring core competencies to recognize and manage emotions, set and achieve goals, appreciate the perspectives of others, establish and maintain positive relationships, make responsible decisions and handle interpersonal situations constructively."[10] It includes understanding both one's own thoughts, emotions and social positions, as well as those of others. The competencies encapsulated in SEL include responsible decision-making, self-awareness, self-management, social awareness and relationship skills.

Harvard University's EASEL lab[11] can be used to classify the domains that fall within it. The possible domains include cognitive, emotion, social, values, perspectives and identity. For this project, we are interested in the non-academic skills most relevant for children affected by disasters, conflict and chronic adversities – namely social and emotional learning.

In emergency and protracted crisis settings, children's ability to learn can be hampered by stress and trauma. Experiences of violence, trauma and extreme stress can affect the brains of children and adolescents differently than adults. The way that children interact with each other, and with adults and authority figures, can be affected by such adverse experiences. This also impacts learning behaviours and working memory. Children can also develop strategies to cope with trauma that can harm their ability to learn and increase the risk of dropout.

These are not learning outcome domains, but they are domains that may be important to measure to understand what is happening with education settings. They are measured in similar ways to other SEL domains.

FIGURE 6. HARVARD'S EASEL LAB DESCRIPTIONS OF EMOTIONAL AND SOCIAL DOMAINS

Domains of social and emotional learning

Social domain: The social domain includes skills that help you accurately interpret other people's behavior, effectively navigate social situations and interact positively with others. Skills in this domain are required to work collaboratively, solve social problems, build positive relationships and coexist peacefully with others.

Specific skills: Emotion knowledge and expression, emotion and behavior regulation, and empathy and perspective taking.

Emotion domain: The emotion domain includes skills that help you recognize, express and control your emotions, as well as understand an empathize with others. Skills in this domain are important not only for managing your own feelings and behavior, but also for interacting with and responding to others in pro-social ways.

Specific skills: Understanding social cues, conflict resolution and social problem-solving, and pro-social and cooperative behavior.

10 https://inee.org/collections/mhpss-and-sel#:~:text=Social%20and%20emotional%20learning%20(SEL,and%20handle%20interpersonal%20situations%20constructively
11 http://exploresel.gse.harvard.edu/

2.7 TIMING

KEY POINTS:

- Timing will be determined by the school year, when information is needed, logistical considerations and the purpose of the assessment

The timing of the assessments should take into account:

- When in the school year you would expect the assessed construct to have been learnt/developed (if appropriate)
- When the information is needed for the decision(s)
- When it is administratively possible to administer the test (minimal disruption to schools, most likely to be able to find children at home)

For example, in the Democratic Republic of the Congo, assessments are planned to provide a baseline prior to implementing programme interventions. These need to be implemented quickly – without delaying programme implementation – to make sure that the assessments can serve this purpose.

Similarly, in Ethiopia, a baseline assessment was conducted at the beginning of 2021, ahead of the start of ECW's MYRP activities. There is a desire to link learning assessments for crisis-affected populations with the existing national assessment system. This may require implementing assessments at a similar time in the calendar year to improve comparability. There are also ongoing crises that affect the ability to develop and implement assessments, and administration may require greater flexibility.

It is important to keep in mind that there may be a lengthy delay (several months) between applying the instrument and reporting (i.e. producing a final version of the information required for decision-making) as time is needed for data collection/entry, data cleaning, primary data analysis, secondary data analysis and the presentation of findings. It is important to keep this in mind to ensure the information from the assessment is available within the policy-making or decision-making window.

2.8 CONTEXTUAL CHARACTERISTICS OF INTEREST

KEY POINTS:

- To draw inference and assist decision-making, further details will be gathered in addition to the learning assessments

The choice of background characteristics you may want to collect will depend on the purpose of the assessment and the questions that decision-makers need to address.

When selecting these characteristics, it is important to only collect information that will be useful. This protects students and staff from unnecessary information being gathered and stored, and minimizes the time burden for participants and the costs of data collection, storage and analysis for the survey organization. Try to avoid errors of collecting surplus information by carefully questioning whether each item assists decision-makers in using the data for the intended purposes.

The background characteristics can be used to provide additional information, disaggregate learning outcomes data ,and analyse trends and patterns in learning outcomes.

Table 12 outlines a set of potential background characteristics of interest. However, an almost never-ending list of characteristics exists and the background characteristics that should be collected will relate to the research questions defined in the early stages. In addition to this list, there may be other characteristics that will be highly relevant in a particular emergency setting.

TABLE 12. POTENTIAL BACKGROUND CHARACTERISTICS OF INTEREST

Student characteristics 	☑ Age
	☑ Grade
	☑ Sex
	☑ Access to counselling
	☑ Time required to get to school
	☑ Population status (refugee, IDP, host community, etc.)
	☑ Time in country
	☑ Language spoken at home
	☑ Disability status and type[12]
Classroom characteristics 	☑ Pupil-textbook ratio
	☑ Number of learning hours per week/day
	☑ Pupil-teacher ratio
	☑ Pupil-classroom ratio
	☑ Classroom set-up (rows, groups, corners, etc.)
	☑ Level of interaction between learners and teacher
	☑ Level of interaction amongst learners
Teacher characteristics 	☑ Age
	☑ Sex
	☑ Qualifications
	☑ Teaching experience
	☑ Language spoken at home
	☑ Dominant pedagogical approach applied
	☑ Level of motivation
	☑ Timely renumerated

12 ECW supports the use of the Child Functioning Module to identify functional disabilities amongst children and adolescents: https://www.washingtongroup-disability.com/question-sets/wg-unicef-child-functioning-module-cfm/

School characteristics	Region/district
	Number of students (girls as a % of total)
	Number of teachers (female as a % of total)
	Number of classrooms
	Number of toilets/latrines, disaggregated by sex
	When it was founded
	Formal/non-formal
	Type (government, private, religious, community, etc.)
	Performance in national examinations.
Household characteristics	Number of adult members of the household
	Number of children
	Employment status of parents
	Income/expenditure of household
	Socioeconomic status
	Reading/learning materials at home

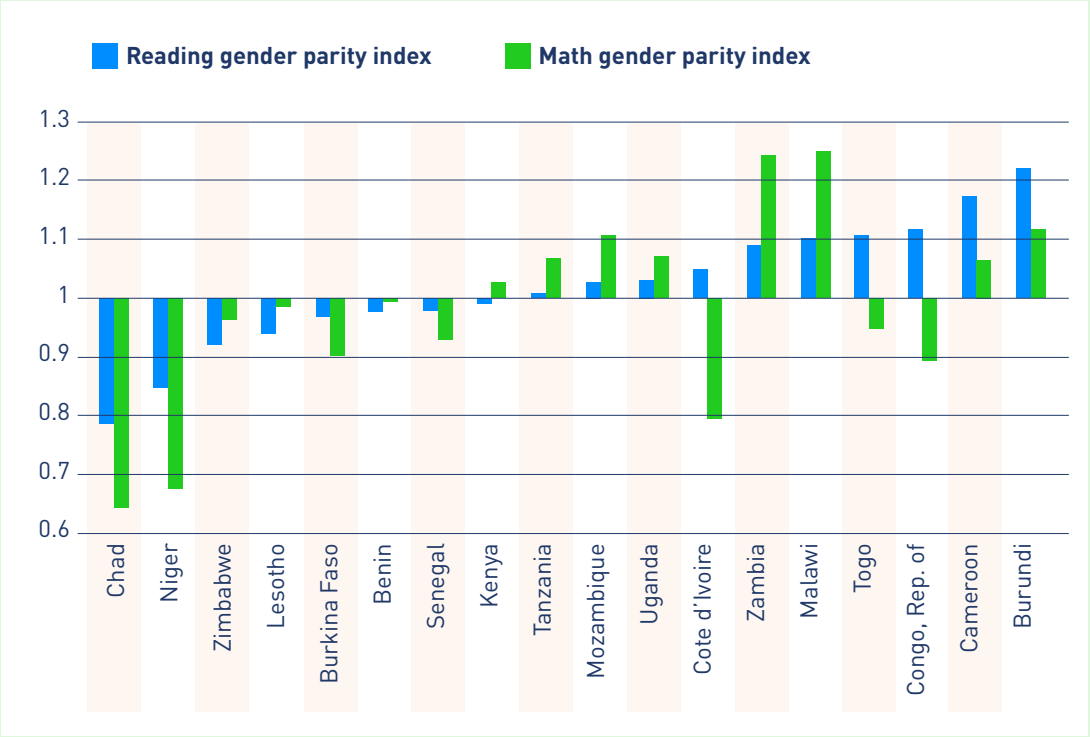
To make sure that useful information is gathered, it is important that the intended users are fully engaged in the process of setting the purposes for the assessments, and in considering the information that they need to do that. Reflection on what would have been helpful from previous learning assessments with a similar purpose (although not necessarily the same target group) may help. Viewing survey examples from similar settings may also help draw up a list of possibilities, which can then be appraised to assess their usefulness for intended purposes.

Disaggregating learning outcomes by gender

One of the most common reasons we collect background characteristics is to disaggregate learning between different kinds of children.

The most common example of this is disaggregating learning by gender. Taken from the Global Partnership for Education (GPE) website, the graph below shows that, on average across 18 GPE partner countries, girls' and boys' performance varies greatly from country to country. We can see the disparity between girls' and boys' performance because, when the learning outcome data was collected, the gender of the student was also collected so that the learning date could be disaggregated by gender. The same can be done for region, urban/rural, etc.

This is a very simple example of how background characteristics can be analyzed, and is called descriptive statistics. More complex analysis are also possible and are reported under Section 3.1.1 Analysis.



2.9 EVIDENCE USE AND UPTAKE

KEY POINTS:

- It is important to identify the primary and secondary users of the information provided by the learning assessment
- When users need the information and the best format for each user should be considered early
- The uptake of evidence should be planned for from the beginning of the assessment process

The 'theory' in that the pathway from measurement to progress towards Sustainable Development Goal 4 should not assume that change happens automatically.

The greatest challenge faced is the utilisation of information generated by learning assessments to drive policy or implementation changes.

Evidence on learning assessments shows that technical soundness is just one of many factors that support the take-up of evidence.[13] Other important factors include timely dissemination of results, the integration of the assessment into existing structures, the involvement of policy and/or decision-makers in the design and implementation of the assessment, the identification of factors associated with high and low achievement, and the findings being presented in a manner which is easily understood by policy makers. From early partnerships such as PARIS21, to more recent assessment programmes, the evidence is clear that successes are most likely when the Paris Declaration on Aid Effectiveness principles are followed – in particular, ownership.[14] Therefore, the success of the programme is not only dependent on building capacity in the technical skills required to carry out the functions of the project, but also on building ownership over the processes and outputs.

In the pursuit of evidence uptake, we are increasingly realising the importance of intelligent content concepts. Drawing from the field of marketing, we now know the importance of:

The right **person**
The right **content**
At the right **place**
At the right **time**
In the right **format**
In the right **language**

Identify the primary users of the information provided by this learning assessment. Be as specific as possible in identifying people within organizations, as well as the organizations more broadly. Next, identify how they will use the information: for which decisions and when? Will the information be used in conjunction with other data?

Finally, identify the kinds of content each stakeholder group will be interested in, how it should be delivered, when it should reach the stakeholder, and what format and language will be most appropriate for maximum likelihood of consumption.

Also, identify other stakeholders who should be made aware of the findings and are also likely to find them useful. Again, be as specific as possible. How would you expect them to use the information generated?

13 Kelleghan, Greaney & Murray (2009)
14 This is not only applicable to government stakeholders, but also other programme architects and decision-makers.

Table 13 provides a series of questions to be answered for each organization or unit receiving information on the results. Filling this in during the design phase will enable stakeholders to plan for dissemination to increase uptake.

TABLE 13. DISSEMINATION PLAN

Organization			
Level of importance (primary/ secondary)			
Who is the right individual?			
How will they use the information?			
What information do they need?			
Where should it be presented?			
When do they need it?			
What format will be most effective?			
What language should be used?			

3
DESIGN AND EXECUTION

Once it is agreed that the proposed design outline meets the requirements of the programme, the GTP and the ACP will go through the detailed aspects of the design process. This will include reviewing what is in place, what further work needs to be done, what information will be elicited, exactly who the assessment will be administered to, how fieldwork will be designed and managed, and what analysis, reporting and dissemination will need to take place.

FIGURE 7. SUMMARY OF DESIGN PROCESS

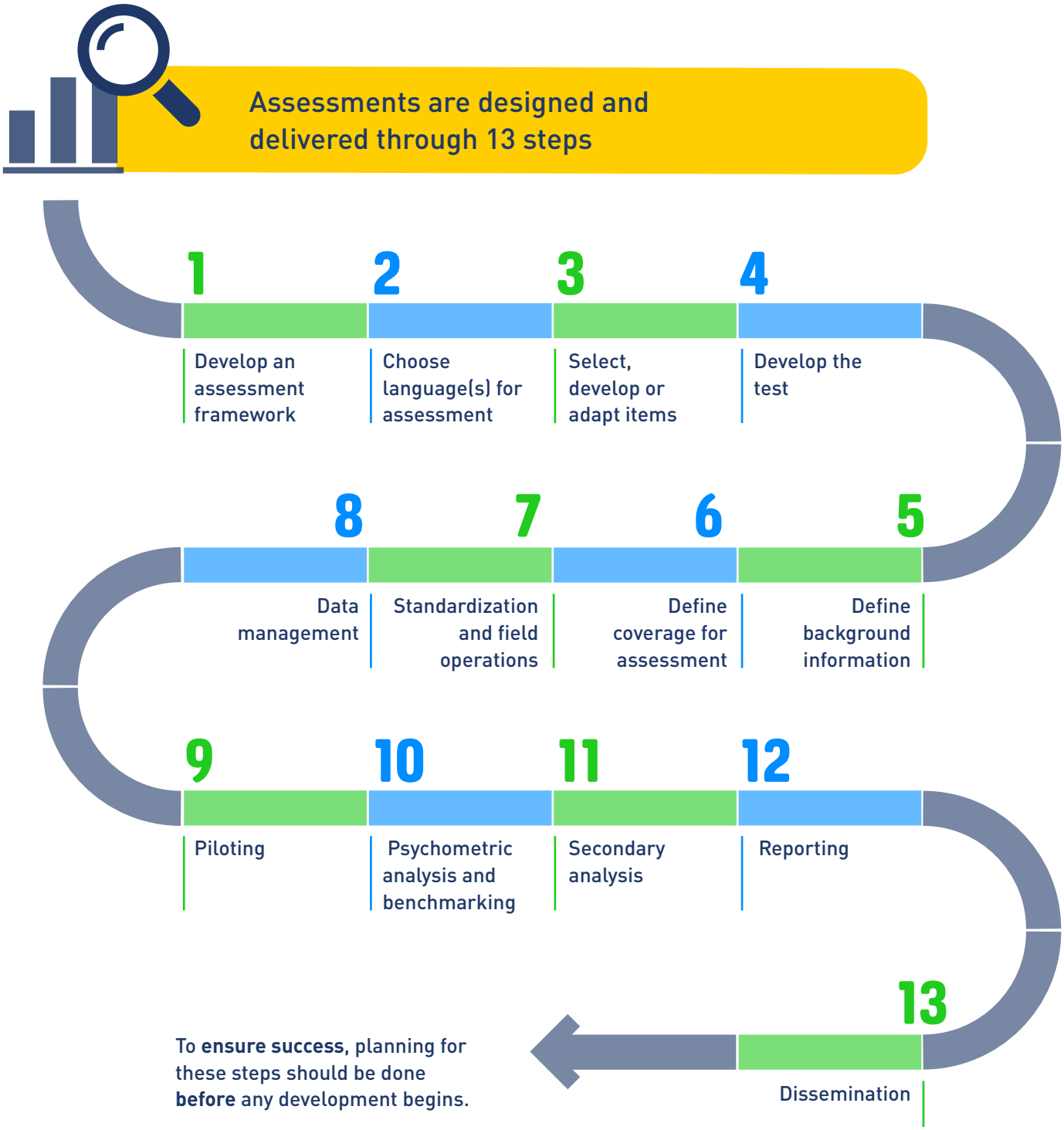


Table 14 lays out the information required during this phase. Details are provided below.

TABLE 14. DESIGN PROCESS QUESTIONS

Assessment areas	Summative assessment	Responses
Assessment framework	Is there an assessment framework in place which aligns to the proposed design of the assessment project/programme?	
Language	What language(s) will the assessment/ assessment training (for formative) be in and why?	
Items	Are items adapted to the context, language(s) of assessment, and do they capture the variety of ability levels of the learners?	
Test development	Does the combination of available items cover (in terms of scope and balance) the assessment domain?	
Background information	What information is needed about students/ classrooms/schools, and how will it be collected and linked to assessments?	
Coverage	Who will the assessment be administered to?	
Standardization and field operations	How will be assessment be standardized across the implementation area?	
Data management	How will data be collected and managed to ensure consistency and quality?	
Piloting	Will the assessment be piloted to ensure quality items and good targeting?	
Psychometric analysis	Will the assessment be reviewed for item and test performance using Item Response Theory (IRT), and will the test be scaled using psychometric techniques?	
Secondary analysis	What analysis will be undertaken?	
Reporting	What will the report cover? Does it still align with the policy/programme goal?	
Dissemination	Who needs to know what, and when?	

ROLES AND RESPONSIBILITIES

It is useful to record who is responsible for all tasks in the process and who else is involved, in order to engage the right people and manage the whole process. Table 17 provides an example framework for doing this.

TABLE 15. MAIN RESPONSIBILITIES

Responsibility		Lead	Key contributors/ consultees
Assessment framework	☑ Identifying the assessment framework ☑ Developing an assessment framework, if none is available		
Language	☑ Determining languages used ☑ Translating items/instruments		
Items	☑ Identifying and gathering available items, together with supporting data ☑ Reviewing available items		
Test development	☑ Mapping test items against the assessment domain, considering difficulty of each item ☑ Developing additional items, as required		
Background information	☑ Determining what background information needs to be collected ☑ Developing instruments to collect background information		
Coverage	☑ Developing sample frame ☑ Developing sampling approach and drawing sample		

Responsibility		Lead	Key contributors/ consultees
Standardization and field operations	☑ Developing field operations manual(s)		
	☑ Manual should include details of who is responsible for:		
	☑ Overseeing field operations and data collection		
	☑ Overseeing child protection procedures		
	☑ Reporting and responding to child protection concerns		
	☑ Maintaining a risk register		
	☑ Organizing travel and accommodation for field staff		
	☑ Communicating with officials		
	☑ Communicating with schools/households/ participants		
	☑ Leading data collection teams		
	☑ Conducting sampling in the field		
	☑ Administering background information instruments		
	☑ Administering learning assessment instruments		
	☑ Scoring learning assessments		
	☑ Entering data from learning assessments (may be part of administration)		
	☑ Transferring data		
	☑ Quality assurance		
	☑ Backing up data		
	☑ Data processing and consolidation		

Responsibility		Lead	Key contributors/ consultees
Piloting	☑ Compiling test items to be piloted		
	☑ Identifying pilot sites		
	☑ Conduct sampling for pilot		
	☑ Plan and communicate timing		
	☑ Conduct fieldwork – preferably with the same team as will be used for the main survey		
	☑ Analysis		
Psychometric analysis			
Secondary analysis			
Reporting and dissemination			

3.1 ASSESSMENT FRAMEWORK

Is there an assessment framework in place which aligns to the proposed design of the assessment project/programme?

An assessment framework links the stated purpose to the test instruments by mapping learning outcomes conceptually. They break down domains into discrete components and describe levels for each. The relevant assessment frameworks need to be identified to ensure that there are no duplicated efforts and that the assessments are aligned with existing frameworks.

The framework should be developed drawing on a range of informed opinion so that it will be widely accepted. This may include academics, curriculum authorities, government representatives and policy-makers, teachers and school managers.

The TIMSS Mathematics Framework is one example. It describes the three domains included in the assessment: number, geometric shapes and measures. Each domain is assigned a weighting, which is the percentage of the total test score that is derived from questions testing the domain.

The framework describes each domain and divides them into sub-domains. For example, the geometric shapes and measures domain begins with the following:

"We are surrounded by objects of different shapes and sizes, and geometry helps us visualize and understand the relationships between shapes and sizes. This topic area deals with understanding measurements, the coordinate plan, lines and angles. It also covers surfaces and solids."

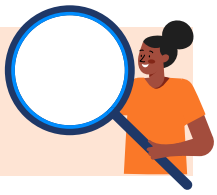
The two topic areas in geometric shapes and measures are as follows:

- Points, lines and angles
- Two- and three-dimensional shapes

It then describes what students should be able to do at the tested grade level:

"At the fourth grade, students should be able to identify the properties and characteristics of lines, angles and a variety of geometric figures, including two- and three-dimensional shapes..."[15]

3.2 LANGUAGES



What language/s will the assessment/assessment training (for formative) be in and why?

Grantees will be aware of sensitivities around language in education in their country. The selection of languages is an important decision and requires a detailed understanding of the language context. There are a few issues to consider when making this decision.

15 TIMSS 2015 *Mathematics Framework*, available at http://timssandpirls.bc.edu/timss2015/downloads/T15_FW_Chap1.pdf

Which languages are required to make sure that students are not prevented from demonstrating their true ability in the selected domain(s)? For non-language domains, if students perform less well in an assessment because of their understanding of the language or ability to communicate in the language, their lower scores will falsely be interpreted as weaker skills in the targeted domain. Therefore, languages should ideally be those that enable students to perform to their optimal level.

Which languages are used in the classroom? It is generally easier to respond to questions in the same language as the child was taught in. Testing in a different language may provide an under-estimate of the child's ability. However, to evaluate the ability of children who have not been in school, it may be fairer to use the home language (in which they have received education previously) or their native language. Selecting a different language that is used in schools would likely result in an over-estimation of the difference between the abilities of students in school and out of school, or between children who have been in the country longer relative to newer entrants.

How many languages do you need? Translating items can be costly as it requires processes to ensure that the translations are reliable and do not significantly alter the way that a question works. Therefore, allowing for the above considerations, the number of languages should be minimised.

The process for translating questions and supporting materials into the languages needs to be designed. Translators will need to be equipped to understand the nuances of the test questions. Ideally, they should be familiar with the subjects being assessed. Translation of test instruments requires a multi-step process. Translations need to be reviewed by experts to make judgements about and improve the quality of the items in the final language.

Selecting an assessment language in Cox's Bazar, Bangladesh

To illustrate the issues, we can consider the situation in Cox's Bazar. People who fled to Bangladesh speak a colloquial language of the Rakhine area (which comes close to the Chittagonian dialect) that does not have any standard written script. The limited number of these children who have had some form of formal schooling in Myanmar have been exposed to Burmese and English – the two official languages of the Myanmar education system. Very few of these children are fluent in either language, most speaking the local Rohingya dialect at home.

The government of Bangladesh directed that both English and Burmese were to be adopted as the medium of instruction in the refugee camps, so that the students would not face any difficulty when they repatriate. In practice, there are multiple languages spoken and taught with a range of proficiency across teachers and students. There are 2 learning facilitators in each centra – one that speaks the local Rohingya dialect, and one from the local community that speaks Bangla. However, the languages taught in the centres are English and Burmese. Very few teachers are proficient in either of those languages.

This context needs to be considered, together with the assessment's purpose to decide on which language(s) to use.

3.3 ITEMS

Are items able to be adapted to the context, language of assessment and the ability levels of the students?

A first step in developing items is to identify what is available to be adapted. These are items that assess the targeted domain. This requires searching for similar instruments that have been used in similar contexts, ideally using the same assessment framework. These will need to be adapted to the context, so they do not have to have been used in exactly the same context with an equivalent group of children. However, they do need to be targeted at a similar level of difficulty (although they may not cover the full range required).

Items should be reviewed for quality at this stage, by looking at how they performed in previous uses and having subject specialists appraise their appropriateness for this use.

As items are collected from other instruments, evidence about their development, uses and psychometric performance (difficulty and discrimination index) should be stored with them in a database. Intellectual property rights should also be observed.

3.4 TEST DEVELOPMENT

Does the combination of available items cover the required scope and balance of the assessment domain?

Is the difficulty of the items aligned to the proficiency of the children being tested?

Tests need to identify differences between test-takers across the whole domain. To do this, we need items that cover the whole domain and the full difficulty range. As a whole, the test must not be too difficult or too easy for the target group.

The available items need to be reviewed to assess whether they cover all the assessment domain for all levels of difficulty required to measure the attainment of the target population.

- First, map each item against the assessment framework. Take note of any gaps in the assessment framework for the assessment domain.
- Then, within each sub-domain, rank items in terms of difficulty, ideally using data from previous uses, but also using expert judgements where necessary. Identify any gaps in difficulty. Consult teachers to check that the questions would not all be too easy or too difficult for the target population.

If new items need to be developed, people with expertise to do so need to be identified. Reviewers will also be required to provide feedback and strengthen items ahead of testing. The reviewers should be close to the test-takers (i.e. teachers) and be able to predict how they might respond. Plan for points during the item development process for feedback.

Initial scoring criteria should be developed as part of the item development process. These should be precise, succinct criteria for each response category. Scoring criteria should be finalised once some evidence as to student responses to the items is available (for example, after a pilot).

3.5 BACKGROUND INFORMATION

What information is needed about students/classrooms/schools, and how will it be collected and linked to assessments?

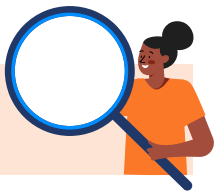
This is a continuation and finalisation of the question in Section 3.8. Finalise the decision about what needs to be collected for the assessment to meet its purpose. Determine which child, school, teacher, classroom and household data is required to enable the analysis that you plan.

Table 12 shows some possible variables that could be collected with learning assessments.

For all the data, identify the most appropriate survey instrument and how it will be administered. The data collection instruments need to provide accurate and reliable information to a level of detail that is sufficient for later analysis. They need to be appropriate to the target population, ensuring that the data gathered is suitable, and that questions take account of the context and participants' ability and comfort to respond. Language should be simple, and instructions must be clear and concise.

3.6 COVERAGE

Who will the assessment be administered to?



Describe the group to be assessed. The ages/grades, school types, geographies, etc. This should include the list of schools and/or teaching centres, the number of students in each and the criteria for selecting students in each site.

The sampling approach needs to be designed with support from a statistician to ensure that results from the assessment provide the required degree of accuracy and validity in interpreting and using the results.

The starting point for the statistical approach is the sample frame, which is the complete population of schools, households, and/or children that you need information about. The sampling frame also needs to include as much relevant information about each unit of the population. 'Relevant information' here is any variables that could be associated (or correlated) with differences in learning outcomes in the chosen domain. The quality of the sample frame is crucial for drawing a sample that delivers robust and valid results that are representative of the age/grade population at the end of the assessment process.

For example, a sampling frame may be a database of all schools in the country. The additional information may include the district, school type, urban/rural classification, number of students, number of teachers, number of classrooms, etc.

If there is a desire to compare specific different groups of students, this will affect the number of students that need to be tested (the more comparisons required, the higher the sample size required). Therefore, it is important to make these decisions early on.

3.7 STANDARDIZATION AND FIELD OPERATIONS

How will the assessment be standardised across the implementation area?

It is important that processes and tests are designed so that:

- All participants are given the same tasks
- All assessments are administered in the same manner
- All assessments are scored in the same way

Therefore, processes of administration and scoring need to be planned so that they are uniform for all participants. Tests are administered by different data collectors and marked by different markers, so processes need to be designed in detail and trained to the same degree of detail.

Consider the ways in which administration could differ between participants, schools and data collectors. Then, go over that list and devise processes to standardize across all participants. Consider timing, materials and the setting. How long should participants be given, what materials do they need, when and how should the materials be administered? Also, consider what can happen that might not be planned. What should happen if the test is interrupted? What should happen if a student is unable to continue?

The processes will need to be formalised in a field operations handbook that will be developed by the AP with support from the GTP. This will describe the procedures and protocols for all stages of data collection, marking and data management. It will also include procedures for what to do should foreseeable problems arise. The field operations handbook should be detailed, including timings and who is responsible for each activity.

Administering learning assessments entails working with children, so safeguarding procedures need to be designed into the administration protocol.

A pilot should then be undertaken, whereby both the items and procedures are piloted. The GTP can work with the AP to decide on the sample size for the pilot, assist with procedures for documenting what did and did not work, and make the appropriate changes.

Costs need to be considered and budgeted for throughout the process. As the field operations component is the most expensive stage, this is especially true for this element. Different fieldwork models should be considered regarding how large each data collection team is, how long each team should spend at each school/household and how teams are managed.

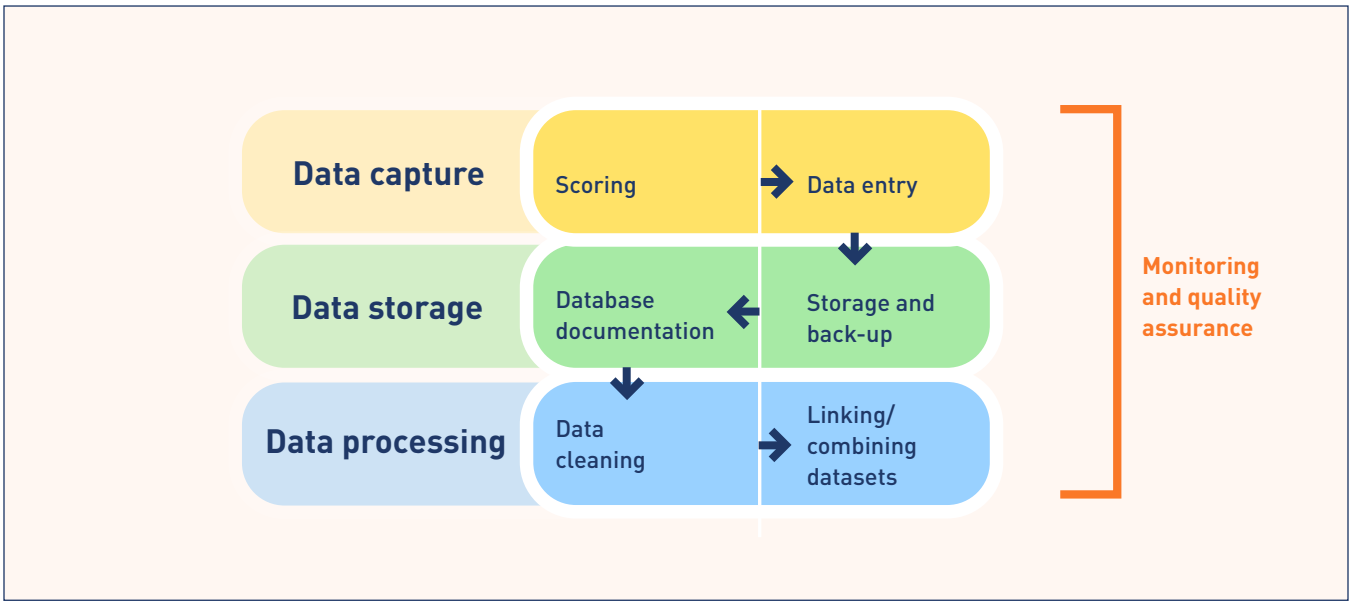
3.8 DATA MANAGEMENT

How will data be collected and managed to ensure consistency and quality?

As a continuation of the decisions above, processes for data entry and marking should also be designed at this stage. The aim is to ensure that data is entered in a uniform manner and that marking and scoring is the same for all participants, schools and data collectors. The end result from this phase is a final database that is ready for analysis, formatted/cleaned as required and free from errors.

The process should include capturing, storing and processing. For each part of the process, timings, roles and responsibilities should be clearly identified.

FIGURE 8. PROCESSES TO GO FROM DATA GATHERING TO THE FINAL DATASET



3.8.1 Data capture

Data capture includes data entry, scoring and monitoring, and quality assurance. Data capture can be paper-based or electronic (Computer Assisted Personal Interviewing/CAPI). It can also utilise electronic data processing such as optical mark recognition. There need to be clear coding rules for all possible responses. This needs to be planned so that all information needed for all indicators is captured.

Scoring is the method of translating a student's response to a question into a score that can be analysed. The scoring method depends on the question and should be designed together with the question. Scoring for multiple choice or short answer questions will likely have one answer identified as 'correct' or 'best', but there may also be partial credit (lower scores) for other answers. Long answer questions require a marking rubric with clearly articulated criteria. Scoring can be automated, recorded by enumerators or conducted by markers after data collection. There needs to be clear training to ensure uniformity across all markers.

Monitoring and quality assurance processes are important to maintain consistent standards across all data collectors and all sites. These may include on-site monitoring and processes to prevent or catch data entry errors. This may help flag issues that need to be addressed by re-visiting test sites. Quality assurance should cover all aspects of data collection.

3.8.2 Data storage

All data (including digital and paper) needs to be stored securely and backed up. Data should be stored with the information needed for analysis and without identifying information. Names of students should be replaced by unique identifiers. The list linking student names with unique identifiers should only be accessible to a very small group of people who have been trained in data security and ethics.

Identify where data is going to be stored, how it is going to be transferred there and where it is backed up. There need to be protocols to consolidate and back up data according to a planned schedule.

The database should be properly documented so that all variables are properly labelled and described. The source, processing undertaken and coding details should all be listed within a codebook.

Where data contains identifiable information, such as names, school names or date of birth, plans should be made to ensure that the data is stored securely.

3.8.3 Data processing

Data needs to be cleaned to ensure that all data is coded correctly, entered completely, and that duplicates and irrelevant data are removed.

Link variables need to be generated and cross-checked if data is going to be linked. Datasets may be combined to provide a consolidated database.

Processing should ensure that the children involved are protected. The data should be anonymised so that it is not possible for anyone to identify who each line of data relates to from the information in the dataset. This starts by only collecting information that is needed. However, anonymisation at the processing stage will likely also be necessary.

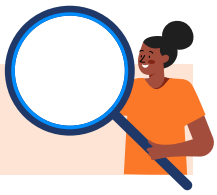
Assess which data might be identifiable and plan ways to anonymise it. Some information is sufficient on its own to identify the individual. This includes names, contact details and images. Others may be combined to enable someone to identify an individual. The risk of identification is greater where there are small categories. That is, very few people share the characteristic recorded. This includes school name and grade, which narrows down the possible children significantly. Other details may include rare conditions or detailed personal histories.

Dates risk enabling identification. These include date of birth and date of entry into the country. These can be anonymised by entering categories (months or quarters) instead of specific dates.

Anonymous identifiers can be used to link datasets without risking identification of individual children.

Overall, the key for anonymisation is to ensure that there is a plan, and that named people are responsible for implementing it reliably.

3.9 PILOTING



Will the assessment be piloted to ensure quality items and good targeting?

Piloting may not be required if items have been used in a very similar setting and with a similar cohort. In these cases, the data from previous uses may be sufficient to demonstrate that the items work as intended. However, if there are new items or significant adaptations, they should be piloted. This is an important step to ensure that tests will provide the information required by working as intended and being targeted at the correct level of difficulty for the target population.

The pilot should be planned with reference to Table 15.

TABLE 16. PILOT DESIGN QUESTIONS

QUESTION	GUIDANCE
Which items are being piloted?	Items that have been newly developed, significantly adapted, or not used in similar settings need to be piloted. It is good practice to pilot two to three times as many items as are needed for the final instrument. This provides a better chance of having enough items that work well enough for the final instrument. Consider how many instruments will be needed to administer the items.
Where will the items be piloted?	Identify the schools, learning centres or communities where the items will be piloted. Ensure that the characteristics and range of ability will be similar to those of the population that will take the final assessment. Do not pilot items with individuals who will be administered the assessment in the actual survey.
How many children will take the tests?	Determine the number children who will answer each test item. This will need input from assessment experts/psychometricians to make sure that numbers are sufficient to provide enough information about item characteristics.
Who will conduct the pilot?	It is best to use the team that will administer the test so that field operations can be practiced and tested at the same time.
When will the pilot be conducted?	This will most likely depend on the timelines required for administration of the final assessment.
What analysis will be conducted?	Plan the analyses of pilot data that will be required. These often include: item difficulty (% of answers that are correct); discrimination (the ability of an item to differentiate among students on the basis of how well they know the material being tested); common mistakes and misunderstandings; key test statistics (mean scores and reliability – Cronbach's Alpha).

3.10 PSYCHOMETRIC ANALYSIS

Will the assessment be reviewed for item and test performance using Item Response Theory (IRT) and will the test be scaled using IRT?

There are a range of psychometric techniques that can be undertaken in order to better understand the quality, validity, reliability and appropriateness of assessments.

- **Validity is an important concept in learning assessment.**

  "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests."[16]

- **Reliability is related to validity.**

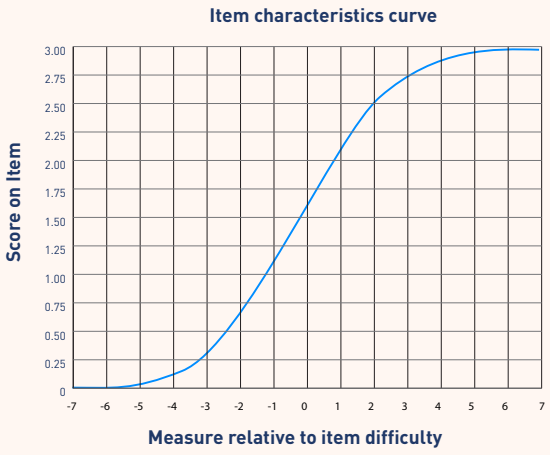
  "Reliability refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups."[13]

Analysis includes:

- **Analysis of test reliability statistics.** This provides information on the reliability of the tools and the extent to which they provide reliable measures.
- **Analysis of item discrimination.** Item discrimination is used to determine how well an item is able to discriminate between good and poor students. A value of over .5 is considered to discriminate well, and a value of over .3 is 'fair'.
- **The production of Item Characteristic Curves.** This provides information on the probability of a student answering the item correctly by their overall score. The probability of answering an individual item should increase as students' overall scores increase.

Item characteristic curve (ICC) is a plot of the probability that a test item is answered correctly against the examinee's underlying ability on the trait being measured.

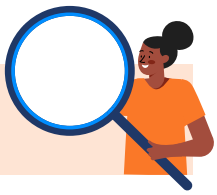
The ICC below is an example of an item that discriminates well. As the student's overall proficiency increases, the probability of the student answering the item correctly increases.



Plan whether Item Response Theory will be used to check that the items and the test as a whole performs properly. Also plan whether it will be used to build a scale so that student performance can be analysed and compared.

16 AERA, APA, NCME (2014) *Standards for Educational and Psychological Testing*

3.11 BENCHMARKING



How will the data be benchmarked against expectations?

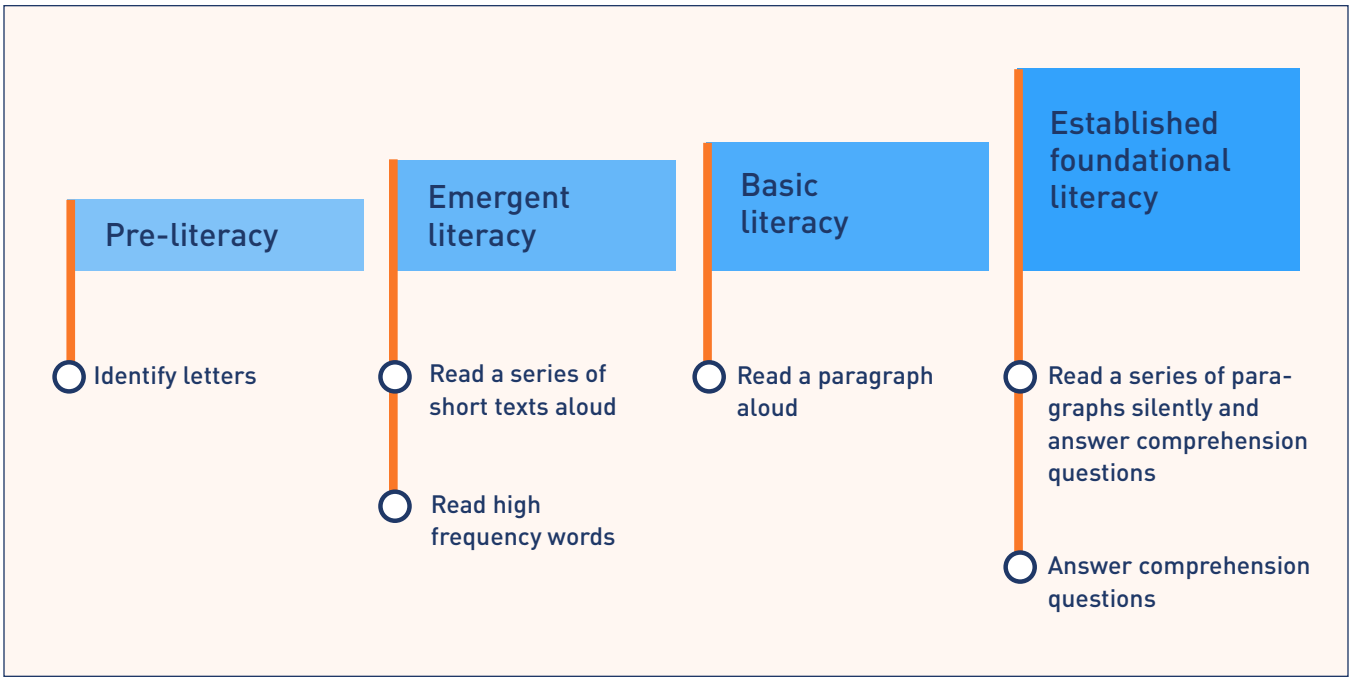
Benchmarking takes scale scores generated by the psychometric analysis and draws cut points to create performance bands. Performance or proficiency bands can be against universal terminology like "Preliterate", "Emergent Literacy", "Basic Literacy" and "Established Foundational Literacy". Where a curriculum exists and is applied in supported schools, the proficiency bands could be against grade level expectations.

The GTP should conduct a workshop to facilitate strategic partners to make decisions such as:

- Whether universal or grade level expectations will be used for benchmarking
- Which knowledge and skills fall within each proficiency level

On the basis of this, the GTP can determine where cut points should be drawn against scaled scores. Figure 9 outlines the four levels and which competencies children have within each proficiency level.

FIGURE 9. EXAMPLE SKILLS AGAINST PROFICIENCY LEVELS



3.12 SECONDARY ANALYSIS

What analysis will be undertaken?

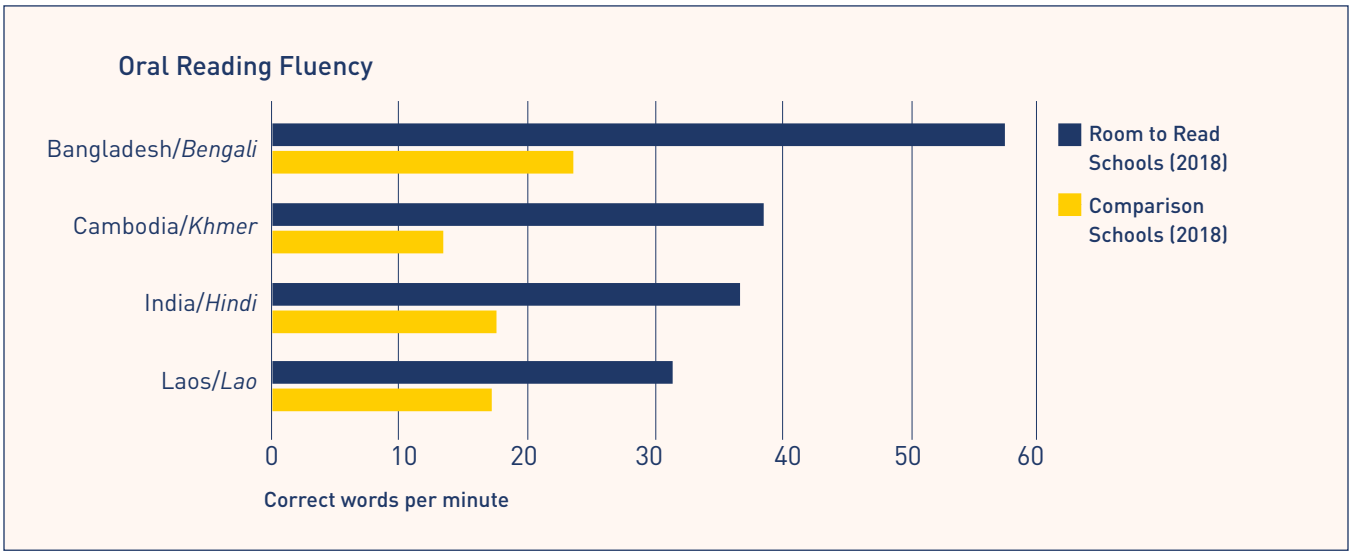
Secondary analysis is any analysis that will be undertaken once the raw test scores have been translated into a usable scale and into performance bands.

The starting point for determining which analysis should be undertaken is to look at the intended purposes of the test and the associated research questions.

The analysis that will be undertaken to support the assessment's intended uses should be planned. Describe the metrics that will be calculated, the comparisons that will be made and the statistical methods used.

One example would be to look at the differences between children who have received an intervention (or set of interventions) and those who have not. Figure 10 shows an example from Room to Read. To provide these comparisons, the method and sample used need to be designed to take account of ('control for') situational difference between children, besides the intervention packages they receive. Statistical analysis would accompany the graph. In this case, it shows that the Room to Read programme was successful in improving reading fluency (on the basis that the comparison schools are similar in all important ways).

FIGURE 10. ORAL READING FLUENCY IN ROOM TO READ AND COMPARISON SCHOOLS, 2018[17]



17 Room to Read (2018) *The End of Impossible: 2018 Annual Report*. Available at www.roomtoread.org/media/bf5fayek/rtr_annualreport_2018.pdf

The easiest approach to analysis is to undertake the analysis by research question.

Research question	Analysis
What are the characteristics of the learners in the supported schools/ centres?	Descriptive statistics of contextual variables
What are the learning levels of Grade 2 students and non-formal students in English and Math?	Proportion of learners achieving at each level of the English proficiency bands Proportion of learners achieving at each level of the math proficiency bands
What are the learners' achievements in social-emotional learning?	Proportion of learners achieving at each level of the social-emotional learning proficiency bands
Are there differences in learning by population status?	Proportion of learners achieving at each level of the English proficiency bands, by population status Average English scale score, by population status Proportion of learners achieving at each level of the math proficiency bands, by population status Average math scale score, by population status Proportion of learners achieving at each level of the social-emotional learning proficiency bands, by population status Average social-emotional learning scale score, by population status
Are there differences in learning by gender?	Proportion of learners achieving at each level of the English proficiency bands, by gender Average English scale score, by gender Proportion of learners achieving at each level of the math proficiency bands, by gender Average math scale score, by gender Proportion of learners achieving at each level of the social-emotional learning proficiency bands, by gender Average social-emotional learning scale score, by gender
What are the important contextual factors associated with difference in learning levels?	Regression analysis including student and home, teacher, classroom and school variables as the independent variables and learning outcomes as the dependent variable. Of particular interest will be characteristics that are open to intervention. For example, teacher qualifications, teacher training, school infrastructure, etc. The GTP can support identifying the kinds of analysis that will be possible and appropriate to meet the policy goal.

3.13 REPORTING

What will the report cover? Does it still align with the policy/programme goal?

Plan the reports and briefs that will be produced. Again, start with the purpose that was defined at the start of the process. Also consider the primary audience for the reports – who will be using the information and making decisions?

Taking these things into account, what information needs to be presented and how should it be presented? List what the report will cover, considering what information should be provided about each of the following:

- Purpose of the test
- Test development process
- Test administration
- Limitations of the test
- Findings
- Conclusions and recommendations

The report should include everything that the users need to know to fulfil the test's purpose, but they should not include too much surplus information that makes it more difficult for users to find what they need.

When presenting the results of an assessment, there are three rules that should always be considered when planning how findings will be presented:

1. **So what?** What do these results mean? A scaled score, when presented on its own, is meaningless without context. For different audiences the 'so what' may be different but, at its essence, the question will be: what does this say about what children know and can do in this country/region/group? Ensure that the results are explained in a way that tells readers what they mean in terms of what children have achieved, and what they need to learn next or more.

2. **From where?** What does good look like for these students? Even if you explain what the scores mean in terms of what children know and can do, you should also consider portraying them in terms of what children should know and should be able to do. The mode and nature of this will vary, but can be:

- **A comparison over time** (i.e. are students improving over time). For this, multiple points of comparable data are needed for meaningful comparison.
- **A comparison to curriculum standards or benchmarks** (i.e. are students reaching the level we have decided is desirable for them). For this, it is important to ensure that whatever standards are used, whether national or international (e.g. SDG 4.1.1/4.1.2), benchmarks are agreed upon by key stakeholders.
- **A comparison to other groups** (i.e. are students here performing at the same level as in other countries). For this, one should be wary of the political implications and usefulness of comparison to other countries, particularly those with differing education systems, or social/economic conditions.

3 **But why?** Why did students perform as they did? Parents, teachers, school leaders and policy-makers, should they engage with results, will all want to know why students performed well or poorly. This question is crucial for beginning to think about policy solutions. While assessment data will never be able to conclusively explain why results were achieved, secondary analysis and the collection of background and contextual data can allow for reporting on what factors influenced the achievement (or non-achievement) of results for different groups, providing insights into questions of why results were as they were.

It is important to remember that once a report is written and has been disseminated, the narrative around results will be controlled by public perception. How results are reported should be carefully thought through with this in mind.

3.14 DISSEMINATION

Who should receive the information when, and in what format?

The main users should already have been identified. At this stage, you will need to finalise this list of users – everyone who could make use of the data. Consult with stakeholders to make sure that it is as comprehensive as it can be. Wherever possible, individuals should be named within organizations. Some users will be particularly prioritised based on the greater impact of their uses of the data. These are the users for which the report structure is particularly tailored, and the dissemination activities will focus on.

For each user, identify in as much detail as possible what they need to know to be able to use the data. When do they need to make decisions and how far ahead of this do they need the information? How does this data interact with other data?

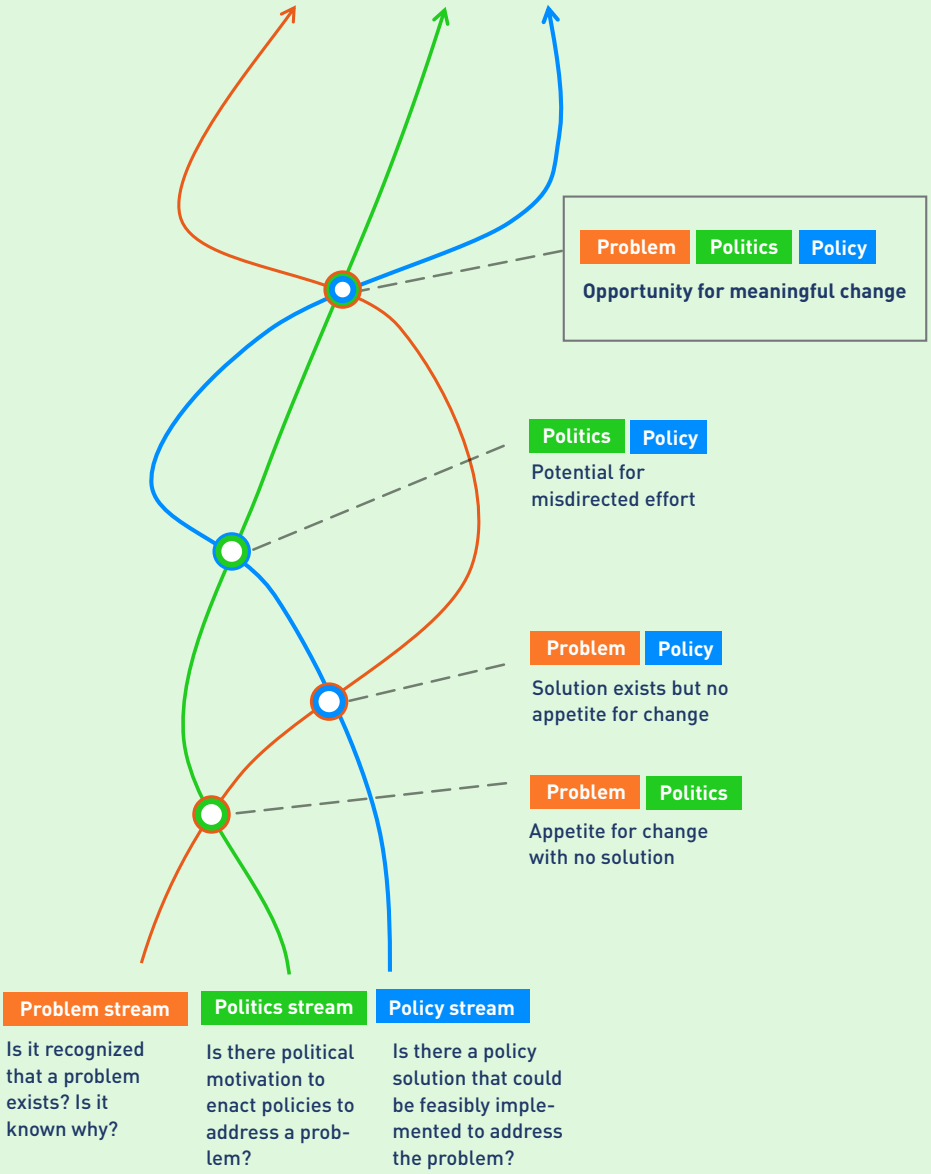
The following table can be used to target specific organizations for research outputs that will arrive at the right time and in the right format.

TABLE 17. DISSEMINATION PLANNING

Organization		
Individuals		
Contact details (email/phone number)		
How will they use the information?		
Primary user? (Yes/No)		
When is the data required?		
How can information be presented to assist proper use?		

Understanding policy change: Kingdon's multiple streams approach

One approach to understanding how policy change occurs, and how it can be supported, is through Kingdon's multiple streams approach. This envisages three separate streams, **problem, politics** and **policy**, with meaningful change only occurring when all three converge. The results of an assessment are most likely to influence the recognition of a **problem**. The approach to dissemination is essential to maximize the impact on both **politics**, as well as encouraging dialogue and further analysis on **policy** solutions.



A. Population sample data template

Implementing partner	Governorate	District	Name of school	Level of school (pre-primary, primary/ secondary)	Total number of grades	Grade level (1-12) *please add lines, as many as needed to show all grades*	Number of classrooms per grade

Number of boys per level	Number of girls per level	Total number of children	% of student who are girls	Number of children with disabilities – girls in formal	Number of children with disabilities – boys in formal	Total children with disabilities	Community composition *Refugee/IDP/ host community/ returnees*

B. Key concepts – validity, reliability and fairness

Validity

Validity is closely related to the concept of fitness for purpose. "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests."[17] It is a judgement that combines empirical evidence with a theory and narrative. Evaluating validity requires examination of the whole process of learning assessment design from item development through to analysis, drawing inference and using the data in decision-making. All elements need to be considered against a clearly defined purpose.

There are two key threats to validity. The first is that factors other than the domain of interest – the skills and knowledge being assessed – affect the scores that test-takers achieve. That is, two children with the same ability in the domain (e.g. reading) could achieve different scores because of other differences between them. This can occur in many ways. For example, if a test uses stimuli that draw on experiences that are familiar to some children and not others, it will likely favour those who are familiar. A test might use examples from farming and agriculture to test literacy skills. This would favour children from rural areas or farming backgrounds for whom the concepts and vocabulary would be more familiar. Alternatively, test questions that closely resemble the style and layout of questions and instruction provided in schools would favour children who have learned in that setting over children who have developed their skills in other countries or in non-school settings.

Outside of test development, other factors that could affect test scores include differences in the settings for tests (children taking them in quieter settings may perform better) or issues with marking (markers may be more generous for children with better handwriting, when this is not part of the construct).

The other key threat is that elements of the domain are not given due prominence within the final test score. Once the domain is defined, the test design and scoring methodology should reflect all sub-domains within and the relative weights given to it. Failure to do so means that the assessment does not measure what it purports to.

Reliability

"Reliability refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups."[13] It answers the question of whether the same result would occur if the assessment was to be repeated. If reliability is poor, then users have less certainty about findings.

As with validity, reliability issues can arise throughout the assessment process. Student performance may vary depending on a range of factors, including the setting for the test, the time of day, the day of the week and details about the question (e.g. subject matter used to test skills like reading and numeracy). Different assessors may act or mark differently and may make different decisions faced with the same student. Where assessments need to be comparable through times, levels and standards may not be consistent across iterations of test instruments.

Fairness

Fairness refers to whether an assessment is free from bias. The test should be appropriate for all respondents, whatever their characteristics (race, religion, gender, age, home language, legal status, etc.). It is important to be mindful of the impact of the assessment on the welfare of the test-taker and take all possible actions to minimize any negative effects. Justification for the assessment should also take account of impact on test-taker.

17 AERA, APA, NCME (2014) *Standards for Educational and Psychological Testing*

C. Framework for answering design questions

Policy Goal	What is the overall goal and research questions of the assessment?	
Resources	What level of resource is available against the ambition?	
Design	The assessment system to be strengthened	*For example: Full integration with national system* *Partial integration with national system* *Non-integration with host or origin system, but with stakeholder alignment*
	School or community based?	
	Target grades or ages?	
	Domains?	
	Timing?	
	Student background characteristics of interest?	
	School and teacher characteristics of interest?	
	Assessment partner?	
	Evidence use and uptake?	

D. Dissemination and uptake tracker

Tracking dissemination and uptake

Reaching the point of dissemination means that partners have successfully brought stakeholders together, agreed on the research purpose and questions of the assessment, implemented a holistic learning outcomes assessment and developed findings, conclusions and recommendations. This isn't an easy task in EiEPC settings! Well-done.

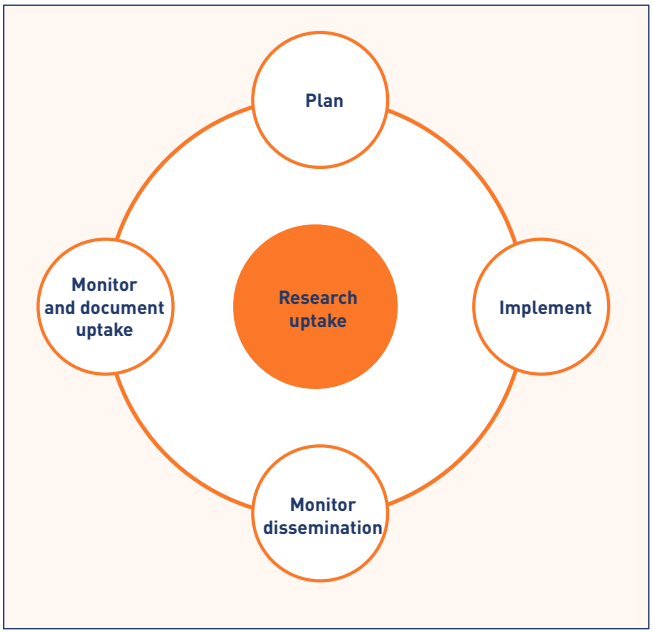
Now is the time to reflect on the design decisions workshop and put into place the dissemination plans developed and track research uptake.

Evidence-uptake in mainstream education settings rarely takes place automatically and is therefore even less likely to take place in EiEPC settings. This points to the importance of purposefully tracking evidence dissemination and documenting uptake. This will increase the chances that the evidence generated will be utilised.

This short template will help you document the dissemination plan, ensure it is implemented in a timely manner, monitor dissemination, and monitor and document uptake.

In this Handbook on designing learning assessment systems, 9 key decisions are outlined to support organizations in developing learning assessment programmes. These questions aim to guide organizations through the process of planning for their learning assessment. These are issues which, if not considered during design, can impact the effectiveness and utility of the eventual learning assessment data. Our 9th decision in our assessment design decisions addressed "how will evidence use and uptake be supported?"

FIGURE 11. THE RESEARCH UPTAKE CYCLE



How will **evidence use and uptake** be supported?

The main users of the evidence should already have been identified at design stage. If needed, you might want to review and adapt your dissemination plan in case priorities have changed.

As the programme grantee, reflecting on the findings of the holistic learning outcomes assessment should be undertaken with implementing partners, and an action plan should be fleshed out for this programme and the next.

Using the table below, the organizations, level of importance, individual(s), information uses, how it should be presented, timeliness, format and language can be identified. This table will determine the outputs for dissemination.

TABLE 18. DISSEMINATION PLANNING

Output number	1	2	3
Organization(s)	*Grantee and implementing partners*		
Level of importance (Primary/ Secondary)	*Primary*		
Who is the right individual?			
How will they use the information?	*To reflect on current and future programme activities and outcomes*		
What information do they need?			
When do they need it?	*Now and in advance of future programme planning*		
What format will be most effective?	*Co-creation workshop reflecting on assessment findings and recommendations*		
What language should be used?			

It is also important to keep track of the development and/or delivery of these outputs, so they are delivered within the decision-making window. The following table can be used to target specific organizations for outputs that will arrive at the right time and in the right format.

TABLE 19. MONITORING IMPLEMENTATION

Output	1	2	3
Date to be delivered/disseminated	*Immediately – 12 December*		
Is support to understand output and use for decision-making needed?	*Yes*		
When and how will this support be provided?	*MEL staff will be in attendance and education advisors*		
What decisions do you expect the user to make based on this output?	*Changes to programme activities based on the findings, covered in an action plan for this programme and the next*		
How will you know if they used the output to make decisions?	*Action plan is implemented*		
What follow-up is needed if the output isn't being used?	*Monitoring IP's changes to activities and designing MEL strategies to monitor effectiveness.*		

Is it also important to track dissemination implementation and monitor if the outputs have been used to inform decisions and assess if any follow-up is needed. This can inform learning, identifying barriers to data use and tracing evidence uptake.

TABLE 20. MONITORING UPTAKE

Output	1	2	3
Disseminated on time and to quality?			
Support provided on time?			
Is there any evidence the output was used to inform decisions?			
Any follow up carried out?			

E. Assessment partner budget template

This template should be provided, in Excel, as part of the request for proposals when procuring the services of an AP.

<NAME OF FIRM>

Offeror should carefully review this template and add any line items they feel necessary to complete the activities outlined in the request for proposals.

Cost heads and items	Unit cost	Persons/ teams/units	Days	Total units	Cost in USD
Labor (illustrative positions, please update according to your proposed plan)					
Engaging with programme grantee and the GTP to design the learning assessment (see the 10 decisions in the terms of reference)					
Team leader	*(Insert daily rate)*				
Assessment specialist	*(Insert daily rate)*				
Other	*(Insert daily rate)*				
Developing an assessment framework					
Team leader					
Other					
Select, develop or adapt items for the assessment and background instrumentation					
Team leader	*(Insert daily rate)*				
Assessment specialist	*(Insert daily rate)*				
Other	*(Insert daily rate)*				
Undertake any required translations and finalize the assessment and background instrumentation					
Team leader	*(Insert daily rate)*				
Other	*(Insert daily rate)*				
Other	*(Insert daily rate)*				
Develop standardization procedures and field operations (including developing enumerator handbook)					
Survey manager/trainer	*(Insert daily rate)*				
Field coordinator/trainer	*(Insert daily rate)*				

Cost heads and items	Unit cost	Persons/ teams/units	Days	Total units	Cost in USD
Programme instruments into STC devices					
Survey manager/trainer	*(Insert daily rate)*				
Field coordinator/trainer	*(Insert daily rate)*				
Train implementing partners and their enumerators on data collection, management and reporting					
Survey manager/trainer	*(Insert daily rate)*				
Field coordinator/trainer	*(Insert daily rate)*				
Pilot instruments					
Team leader	*(Insert daily rate)*				
Survey manager/trainer	*(Insert daily rate)*				
Clean data					
Data manager	*(Insert daily rate)*				
Analyse data					
Team leader	*(Insert daily rate)*				
Other	*(Insert daily rate)*				
Report writing					
Team leader	*(Insert daily rate)*				
Other	*(Insert daily rate)*				
Conduct validation workshop with key stakeholders					
Team leader	*(Insert daily rate)*				
Adapt report					
Team leader	*(Insert daily rate)*				
Other	*(Insert daily rate)*				
Subtotal labor					

Travel and other direct costs					
Training					
Training materials and printing of manuals (per person)					
Training facility rental (rooms/day)					
Transportation to/from training facility, if required (person/day)					
Accommodation, if required for out-of-town enumerators (person/day)					

Cost heads and items	Unit cost	Persons/ teams/units	Days	Total units	Cost in USD
Lunch and coffee breaks during field staff training (person/day)					
Document translations – for training purposes only (if any)					
Other…					
Other…					
Subtotal training					
Pilot administration					
Field staff equipment (bag for tablet, badges, etc; per person)					
Car rental and fuel for team (teams/ day)					
Meal allowance for supervisors and data collectors (per person per day)					
Accommodation for supervisors and data collectors (per person per day)					
Communications allowance (per person)					
Meal allowance for field management (per person per day)					
Accommodation for field management (per person per day)					
Car rental and fuel for field management (per car per day)					
Tablet rental (with GPS and WiFi capability)					
SIM card allowance					
Document translations – for pilot administration purposes only (if any)					
Other…					
Other…					
Subtotal pilot administration					
Survey administration					
Field staff equipment (bag for tablet, badges, etc; per person)					
Car rental and fuel for team (teams/ day)					

Cost heads and items	Unit cost	Persons/ teams/units	Days	Total units	Cost in USD
Meal allowance for supervisors and data collectors (per person per day)					
Accommodation for supervisors and data collectors (per person per day)					
Communications allowance (per person)					
Meal allowance for field management (per person per day)					
Accommodation for field management (per person per day)					
Car rental and fuel for field management (per car per day)					
Tablet rental (with GPS and WiFi capability)					
SIM card allowance					
Document translations – for survey administration purposes only (if any)					
Other...					
Other...					
Subtotal survey administration					
Total cost (not including VAT)					
VAT					
Total cost (including VAT)					

About Education Cannot Wait (ECW):
Education Cannot Wait is the global fund for education in emergencies and protracted crises within the United Nations. We support quality education outcomes for refugee, internally displaced and other crisis-affected girls and boys, so no one is left behind. ECW works through the multilateral system to both increase the speed of responses in crises and connect immediate relief and longer-term interventions through multi-year programming. ECW works in close partnership with governments, public and private donors, UN agencies, civil society organizations, and other humanitarian and development aid actors to increase efficiencies and end siloed responses. ECW urgently appeals to public and private sector donors for expanded support to reach even more vulnerable children and youth.

Additional information is available at
www.educationcannotwait.org
Contact: info@un-ecw.org

Follow us:
@EduCannotWait