
Image Source: Inspiring Teachers

**Practitioner's Guide**

GLOBAL SCHOOLS FORUM

# Measuring foundational learning outcomes: Taking the first steps

## Impact at Scale Labs

April 2026

www.globalschoolsforum.org

GLOBAL SCHOOLS FORUM

## About GSF and Impact at Scale Labs: Foundational Literacy and Numeracy (FLN)

Global Schools Forum (GSF) is a community convenor, knowledge accelerator, funding catalyst, and partnership builder supporting a network of more 200 organisations running or supporting 1.4 million schools and early childhood development (ECD) centres across 70+ low- and middle-income countries. This brief draws on insights from participants in GSF's Impact at Scale Labs on Foundational Literacy and Numeracy (FLN) programme, with whom we work closely to test and refine approaches to improve FLN outcomes in real-world contexts.

Through the Labs, GSF provides organisations with technical support, catalytic funding, and research partnerships to strengthen the evidence base for their FLN programmes. Over 18 months, GSF coaches work closely with each organisation to sharpen their Theory of Change, design testing plans, select appropriate tools and methodologies, and build internal capacity to implement research assessments rigorously. Throughout the monitoring, evaluation, and learning (MEL) journey, GSF coaches support organisations to track progress against research questions, interpret emerging findings, and adapt plans where the evidence required it, ensuring that results were credible and decision-ready. Drawing on this iterative process, the brief shares insights grounded in implementation realities: what organisations choose to measure, the trade-offs they navigate, and the contextual factors shaping their research.

## About the Practitioner's Guide

This Guide focuses specifically on FLN. While foundational skills are typically expected to be developed by the end of Grade 3, many older children continue to acquire them later in their learning journey. For this reason, the assessment approaches outlined here may also be relevant where foundational gaps persist among older learners.

This Guide is intended as a starting point, not a prescription. Organisations working in different contexts will need to adapt these approaches to fit their learners, systems, and resources. Our aim is to bridge assessment theory and field practice by offering the practitioner community a resource grounded in real programme experience, supporting stronger and more evidence-driven approaches to foundational learning.

GLOBAL SCHOOLS FORUM

## Who is this Guide for

This guide is for programme and monitoring, evaluation and learning staff at organisations implementing FLN programmes. It offers practical guidance to support teams whether they are measuring learning outcomes for the first time or strengthening and refining existing assessment practices. The five sections of this guide follow the full measurement journey. This can be used in sequence or start with the section most relevant to your organisation and return to others as your systems and practices develop.

## What this Guide covers



Before selecting any tool, you need to know what question you are trying to answer. We explain how to anchor assessment choices within a Theory of Change- focused on programme-level decisions, so the data you collect is linked to clear goals from the start.



We compare the two most widely used FLN assessment formats: Early Grade Reading and Mathematics Assessments (EGRA and EGMA) and Teaching at the Right Level (TaRL) Assessments. We offer a practical checklist to guide tool selection based on your context, capacity and learning questions. We also cover when to use an existing assessment form, when to adapt one and when to develop something new.



Quality data depends on rigorous planning. We walk through sampling approaches, data collection modes, enumerator training essentials and quality assurance practices that prevent common data integrity failures in the field.



We explain four approaches to analysing FLN data: benchmarks, level-based indicators, zero reduction and mean comparisons, and when each is most useful. Each approach provides a different, complementary view of progress, helping you understand who is improving (for eg., learners starting at lower versus higher levels) and how much progress different groups of students are making.



We outline steps for using assessment data in regular programme review and planning. This will help you use learning data to make concrete decisions about the core intervention, the implementation support required, and what adjustments may be needed to ensure the programme is reaching all learners.

# Introduction

## Why it matters

### Measuring foundational learning outcomes

Improving foundational learning globally is critical. If 90% children could read by age 10, worldwide GDP per capita would be 27% higher by 2050. Hundreds of millions more children would complete primary school, more young people would access stable, meaningful employment, and millions of child deaths could be averted. Yet students who struggle academically in first grade face widening achievement gaps that become increasingly difficult to remediate (GEEAP, 2025).

> Foundational skills refer to basic literacy, numeracy, and transferable skills, that are the building blocks for lifelong learning. - World Bank Blog, 2021

Measuring FLN outcomes represents a commitment to knowing whether children are actually learning, and to using the data to improve programmes. Even when interventions focus on strengthening teacher capacity or overall system delivery, measuring student-level outcomes is the only way to determine whether these interventions lead to learning gains. This also enables investments to be directed toward the approaches that most effectively build the foundational skills that students need to succeed in later grades.

**Measuring foundational learning outcomes is about linking assessments to a clear purpose, choosing quality tools that fit local needs, planning logistics for high-quality data collection, analysing data meaningfully and using findings to inform practice.**

Throughout the guide, examples from Impact at Scale Labs grantees illustrate how the teams navigated different measurement considerations what trade-offs they faced, what they decided and what other organisations can adopt for their own context.

We thank our grantees from the Impact at Scale Labs: Foundational Literacy and Numeracy for their commitment to foundational learning and for contributing their experience to a resource the wider field can use: Angaza Elimu (Kenya), Axium Education, (South Africa), Duara Education (Kenya), Inspiring Teachers (Ghana), Leadership for Equity (India), Peepul (India), and Room to Read (Tanzania).

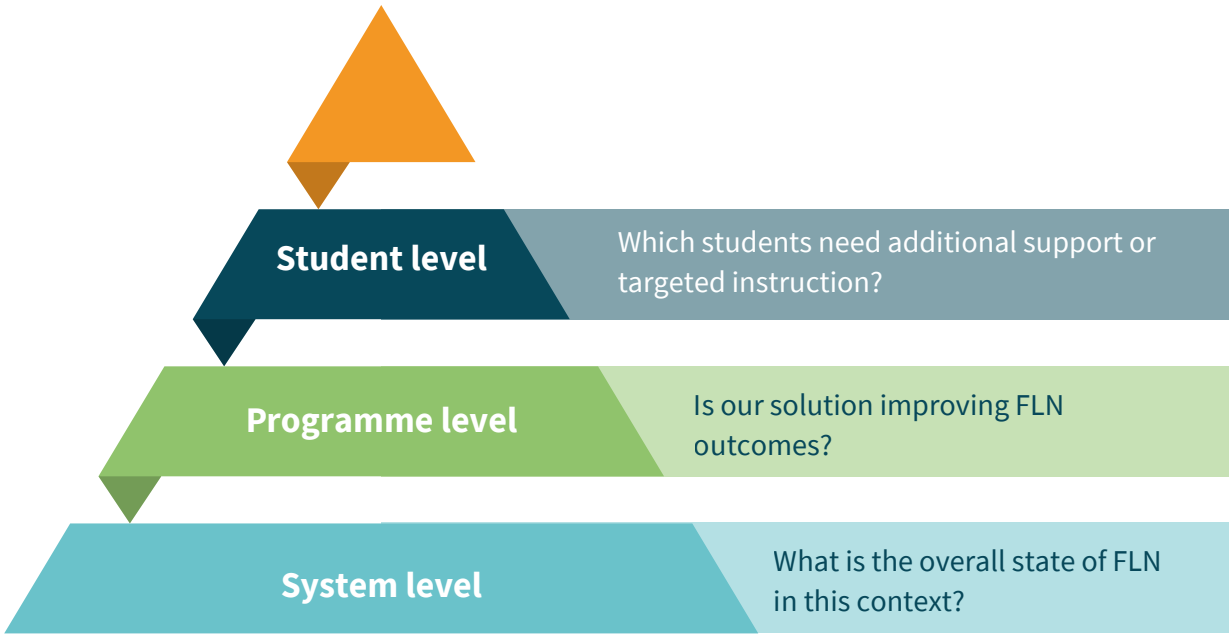
GLOBAL SCHOOLS FORUM

# Five steps to start measuring foundational learning outcomes

| | |
|---|---|
| Define the purpose of assessment | to ensure data collection has a purpose and is linked to a theory of change |
| Select fit-for-purpose tools | that balance validity, reliability, usability and cost in specific context |
| Plan data collection systematically | with clear logistics, training and timelines to ensure data quality |
| Analyse assessment results appropriately | to interpret results meaningfully and track progress over time |
| Connect data to decision making | through structured feedback loops to iterate and improve |

# 1. Define the purpose of assessment

There are several different reasons an organisation may be interested in measuring foundational literacy and numeracy skills. The purpose of measurement shapes both what data is collected and how the results are interpreted. Broadly, learning outcomes can be examined at three levels, each serving a different decision-making need.




Most organisations delivering educational solutions are interested in learning outcomes at the **programme level**, as this data supports practical decisions about programme design, implementation, and improvement. For this reason, the discussion that follows focuses on how different types of foundational learning assessments can be used to answer a programme-level question: **Is our solution improving foundational literacy and numeracy outcomes?** It is important to be clear about why the data is being collected and how it will inform decisions about the programme. This, in turn, requires thinking about assessment choices within a **Theory of Change**.

GLOBAL SCHOOLS FORUM

# Theory of Change

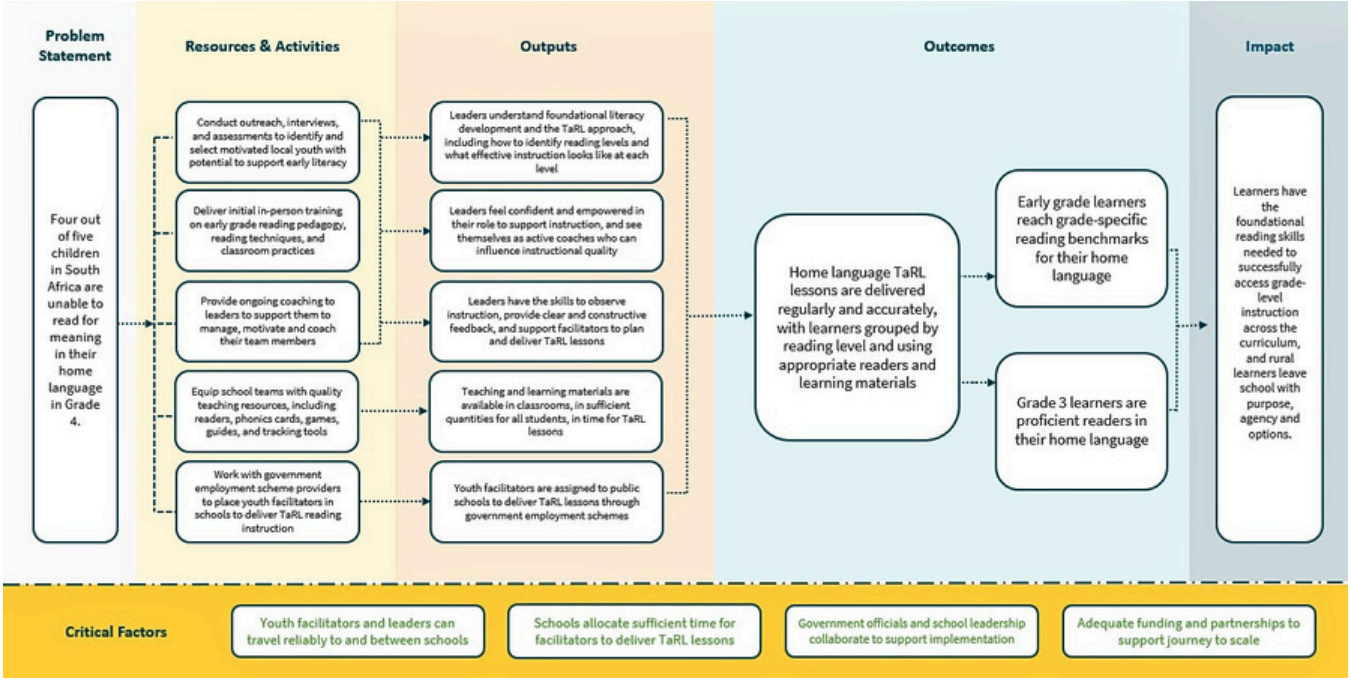
A **Theory of Change** sets out the intended impact pathways leading to improved learning, including the sequence of changes that need to take place along the way. In most foundational learning programmes, improved student learning is the main outcome of interest. However, that improvement depends on a number of steps, including the activities a programme delivers, how those activities are expected to influence teaching and learning practices and which aspects of student learning should change as a result. A clear Theory of Change helps distinguish between these different levels of change and supports more purposeful decisions about what to measure and when.

For support in developing or refining your Theory of Change, see the **Impact at Scale Toolkit: Theory of Change**.

If a programme is designed to influence learning beyond foundational literacy and numeracy—for example, upper-primary reading comprehension or social-emotional skills—then the measurement approach should focus on tools that are aligned with those specific outcomes. However, if work on the programme's Theory of Change shows that foundational literacy and numeracy are among the outcomes the programme expects to influence, the next step is to select assessment tools that can accurately and consistently track progress in these areas.

For additional guidance on aligning measurement choices with outcome pathways, see the **Impact at Scale Toolkit: MEL Framework**.

**Axium Education's Nobalisa Programme Theory of Change,** refined through the Impact at Scale – FLN Labs. It outlines the pathway from programme activities to intermediate outcomes and, ultimately, improved learning in higher grades. Since the Theory of Change anticipates gains in foundational literacy, a foundational literacy assessment is appropriate for programme monitoring and evaluation.




## 2. Select fit-for-purpose tools

In education research, we are often trying to measure concepts like reading comprehension, number knowledge, or empathy—skills and abilities that we cannot observe directly. To measure them, we can ask students direct questions or ask them to perform tasks that give us information about these skills. For example, if we want to understand a student's "number knowledge," we might assess their ability to recognise numbers, identify the larger number, and identify a missing number in a sequence. Together, these tasks give us an understanding of how well students "know" numbers.

### Assessment checklist

When selecting a tool to measure foundational literacy or numeracy, it is important to ensure that the assessment will generate information you can trust and act on. The checklist below summarises key features to look for when choosing a tool.

| Characteristic | Description | Guiding questions |
|---|---|---|
| Standardised | Clear, consistent procedures for administration and scoring so the assessment is delivered the same way each time | • Is there a documented administration protocol, including a script?<br>• Are scoring rules clearly defined? |
| Simple | Able to be administered correctly by enumerators with limited training | • Are the instructions easy for assessors to follow?<br>• After training, do most assessors score the assessments correctly? |
| Sensitive to change | Able to detect a realistic level of progress over time | • Is the assessment appropriate for measuring change at the skill levels of your learners?<br>• Is it able to measure the level of progress that is likely during your implementation period? |
| Relevant | Accounts for the cultural, linguistic, and educational context, including the national curriculum or other relevant frameworks | • Is the assessment aligned with the skills your programme targets?<br>• Is it aligned with the national curriculum or other relevant learning standards?<br>• Is the language of the assessment appropriate for your learners? |
| Feasible | Can be implemented with available budget, staffing, and time | • Is the cost of implementation within your budget?<br>• Do you have sufficient staff to administer it as planned?<br>• Is the time required per child manageable?<br>• Do you have the systems needed for data entry and analysis? |
| Appropriate for pre- or early readers | Designed for learners still developing early reading skills, whose abilities may not be captured by written assessments | • Does the tool include oral tasks where needed? |
| Technically adequate | The assessment measures the skills we care about ("validity") and gives us consistent results ("reliability")* | • Has the tool been validated for your population or a similar one?<br>• Are reliability metrics available and acceptable?<br>• Are there clear methods for interpreting results? |

* See Appendix I for detail on the types of validity and reliability, and how they are defined.

## Assessment formats

Different foundational learning assessment formats vary in characteristics and demand, so selecting the right one depends on what you want to learn and the resources you have available.

The two most common FLN assessment formats are the **Early Grade Reading and Mathematics Assessments (EGRA/EGMA)** and **Teaching at the Right Level (TaRL) Assessments**. Tools using these formats are available in many languages and have all the characteristics of quality assessments described in the assessment checklist. Many other assessments with different names follow the same basic formats, so understanding their core features helps you choose the approach that fits your purpose. EGRA/EGMA and TaRL assessments have been developed and administered extensively in Asia, Africa, and Latin America, so there are no geographic restrictions to using either type of assessment.

**EGRA and EGMA** use multiple short tasks to measure different aspects of foundational reading and mathematics. Although the exact tasks included can vary, EGRA typically includes identifying letter names or sounds, reading real or invented words, and reading a passage and answering basic comprehension questions. EGMA tasks typically include identifying numbers, identifying the larger of two numbers, identifying missing numbers on a number line, basic addition and subtraction problems, and word problems. Many tasks are timed at one minute, allowing these assessments to capture both accuracy and automaticity.

**TaRL Assessments** were first developed by Pratham Education Foundation for use in TaRL programmes and their citizen-led, household-based survey, Annual Status of Education Report (ASER). These assessments are untimed and use a small set of reading and numeracy tasks to identify the highest level a child can complete accurately. In reading, tasks typically progress from letters or syllables to words, a short paragraph, and a simple story. In mathematics, tasks range from number recognition to basic operations.

Additional guidance on how these tools were developed, what they measure and how they are used is available in the **EGRA Toolkit**, **EGMA Toolkit**, and **TaRL Assessments**.

GLOBAL SCHOOLS FORUM

# Choosing a foundational learning assessment type

The following characteristics can help you identify the right assessment tool.

| Characteristic | EGRA and EGMA | TaRL Assessments |
|---|---|---|
| Validity and reliability | High | High |
| Sensitivity to change | Captures small improvements, even at higher levels | Captures larger improvements |
| Level of detail of data | Higher: individual scores for each subtest (e.g., correct words per minute for oral reading fluency), with accuracy and fluency data | Lower: single level on a scale (e.g., proficient at word level), with accuracy data only |
| Enumerator training | Requires relatively more | Requires relatively less |
| Time | Takes longer to administer | Quicker to administer |
| Mode | Individual and oral, recorded digitally or with paper and pencil | Individual and oral, recorded digitally or with paper and pencil |
| Technically adequate | The assessment measures the skills we care about ("validity") and gives us consistent results ("reliability")* | • Has the tool been validated for your population or a similar one?<br>• Are reliability metrics available and acceptable?<br>• Are there clear methods for interpreting results? |

# Illustrative tasks in EGRA/EGMA and TaRL Assessments

EGRA and EGMA include multiple sub-tasks, such as “Number Identification,” shown here. Children complete all sub-tasks and receive a separate score for each one. For example: “identified 10 numbers in one minute.”

**Task 1: Number Identification** | A | 60 seconds

**Here are some numbers. I want you to point to each number and tell me what the number is. I will tell you when to begin and when to stop.** Here are some numbers. I want you to point to each number and tell me what the number is. I will tell you when to begin and when to stop.

- [point to first number] Start here. Are you ready? . . . Start.
- What number is this?

( / ) Incorrect or no response

( ] ) After the last number read

| | | | | |
|---|---|---|---|---|
| 2 | 9 | 0 | 12 | 30 |
| 22 | 45 | 39 | 23 | 48 |
| 91 | 33 | 74 | 87 | 65 |
| 108 | 245 | 587 | 731 | 989 |

- If the time on the stopwatch runs out (60 seconds).
- If a child stops on an item for 5 SECONDS.

TaRL assessments, such as the mathematics example shown here, are organised into levels. Children begin with the easiest level and move up if they complete it successfully. Their score is the highest level they complete. For example: “proficient in subtraction.”

| Number recognition 1&9 | Number recognition 10&99 | Subtraction | Division |
|---|---|---|---|
| 1 4 | 51 83 | 46 − 29 = ____ ; 63 − 39 = ____ | 7 ) 879 |
| 7 3 | 37 65 | 47 − 28 = ____ ; 45 − 17 = ____ | 6 ) 824 |
| 6 9 | 55 26 | 92 − 76 = ____ ; 84 − 57 = ____ | 8 ) 985 |
| 5 2 | 91 43 | 52 − 14 = ____ ; 66 − 48 = ____ | 4 ) 517 |
| | 36 27 | | |

## Adaptations of EGRA, EGMA and TaRL Assessments

There are newer assessment formats with emerging evidence that are also useful to explore. Some build directly on EGRA and EGMA but remove the need for 1:1 administration, such as the Group Administered Literacy Assessment (GALA) and the Self-Administered EGRA and EGMA.

The self-administered EGRA— delivered on a tablet and consisting of multiple-choice items and a spelling task—is a particularly promising tool. Multiple students can be assessed at once, each on their own tablet, removing the need for resource-intensive, 1:1 assessment. In three validation studies conducted in Malawi (Chichewa), Ghana (English), and Kenya (English)1*, scores on the self-administered EGRA and EGMA were highly correlated with scores on the traditional EGRA and EGMA, which are widely considered the "gold standard" for measuring FLN outcomes. This is the first time a group-administered assessment has closely matched the results of these gold-standard tools, suggesting a promising pathway for measuring foundational skills more efficiently at scale.

In addition, PAL Network's International Common Assessment of Numeracy (ICAN) and International Common Assessment of Reading (ICAR) maintain the low cost and easy administration of TaRL Assessments, while fully aligning with the Global Proficiency Framework.

## Selecting test forms

Once you have chosen the type of assessment to use, you can begin identifying or creating specific test forms. Test forms must be well-suited to the context, in language and content. While the standard EGRA, EGMA, and TaRL Assessment formats consistently result in test forms with high validity and reliability, they do not *automatically* do so. It is important to use specific test forms that have been validated to ensure that the content does not privilege one group over another (for example, by requiring knowledge more familiar to urban than rural children), is comparable in difficulty to earlier assessments, and is internally consistent. To do this, you can:

- **Use available forms**: EGRA, EGMA and TaRL Assessment forms have been validated for a number of languages and regions over the past 20+ years. You can find open-source options online (e.g., the Schools 2030 Assessment Bank) or connect with organisations implementing early grade literacy and numeracy programmes in your context. Many organisations, like Save the Children, have developed test forms for their own programmes and may be willing to share them if you reach out directly.
- **Modify existing forms:** There are specific guidelines for adapting previously validated forms, described in assessment-specific resources such as the EGRA Toolkit, EGMA Toolkit, and TaRL Assessment Tool Guidance. These resources explain how forms can be adapted (for example, by changing story characters or reordering items within a row). Adapted forms must still be validated with learners to ensure the changes have not altered the difficulty of the test.

- **Develop new forms**: Partner with external specialists in foundational literacy and numeracy assessment development to create bespoke tools tailored to your specific work and context. New forms must be piloted before administration with the children in your context.

**Case study: Tool selection**

### Peepul's selection of the right tool to measure FLN competencies in middle grades

In Madhya Pradesh's (India) Sandipani Vidyalaya, Peepul recognised a critical challenge: students in Grades 6–8 were often significantly behind grade-level expectations and even lacking in FLN skills. To address this, Peepul supported Middle School Headmasters (MSHMs) in coaching teachers to deliver targeted remedial instruction. Through the Labs, Peepul aimed to measure gains in both foundational literacy and numeracy as well as skills closer to students' grade-level expectations.

Thus, the team needed an assessment tool that covered both foundational skills (basic reading and math) and grade-adjacent competencies (upper primary level tasks). Other considerations included:

- Feasibility for rapid administration at scale,
- Availability in the languages used in the programme (English and Hindi), and
- Possibility of benchmarking results against comparable national data.

A standard ASER tool would have been too easy, leading to ceiling effects that masked real differences between students. Conversely, a grade-level assessment would have been too difficult, with most students scoring zero—providing Peepul with limited insight into actual progress.

Peepul worked through the assessment checklist with GSF and used it to systematically compare fitment of various tools. The checklist helped the team to clarify what they needed the data to do, which populations they were measuring, and what kinds of comparisons would be meaningful.

After careful consideration, Peepul developed an assessment tool combining ASER Basics and ASER Beyond Basics. This tool included the standard ASER tasks for foundational literacy and numeracy with additional task sets designed to measure more advanced competencies, like measurement, reading and understanding instructions, and performing financial calculations.

These additional tasks are aligned with upper-primary learning levels. The tool is also standardised, simple to administer and well-suited to organisations that need to train enumerators efficiently. It is supported by prior evidence of validity, requires minimal resources and can be implemented at low cost. Importantly, the availability of national ASER data meant Peepul could situate their findings against relevant reference points.

| What to measure | Assessment tool | Content |
|---|---|---|
| Foundational learning competencies | Annual Status of Education Report (ASER) | Reading letters, words, paragraph, story; Identifying numbers 1-9 and 10-99; Subtraction (2-digits with borrowing); Division (3-digits by 1-digit) |
| Near grade level competencies | ASER Beyond Basics | Applied tasks - Everyday calculations like calculating time, adding weights, measuring length, word problem involving financial calculation (competencies aligned with grades 3 and 4), advanced comprehension aligned with grade 5 learning outcomes |

Ultimately, the decision was about choosing an assessment that aligned with Peepul's programme logic and learning questions. By pairing a technically sound tool with a clear purpose, the team ensured that their data would be able to answer the critical questions they cared about most, that is, what is the learning gap in middle grades of Sandipani Vidyalaya for students?

GLOBAL SCHOOLS FORUM

# 3. Plan data collection systematically

Once you have chosen or developed your tool, the next step is to plan data collection. The quality of assessment data depends largely on the foresight applied to data collection planning. Even the most technically sound assessment tools may produce unreliable or biased results if the data collection process is not systematically designed, considering contextual realities, logistical constraints and adherence to standardised protocols.

## Key considerations during the pilot or pre-data collection phase

- **Engage with government authorities and schools:** Initiate consultations with departmental officials and school heads at least 4–6 weeks prior to actual data collection phase. Through formal workshops or meetings, align on the assessment objectives, scope (e.g., targeted grades, locations), proposed schedule, required operational and logistical support. Ensure formal documentation is in place, such as issuance of official letters or directives to sampled schools specifying assessment stipulations, expected timings, enumerator responsibilities and parental consent.

- **Align with academic calendars:** Schedule assessments to coincide with optimal learning periods, avoiding conflicts with school holidays, examinations or peak instructional periods. For baseline assessments, target the early academic term (e.g., 4–6 weeks after school reopening); for endline, aim for the final term. If you are just getting started measuring learning outcomes, begin with two assessments per year—one at the beginning and one at the end of the school year. As your capacity, resources and data-use systems develop, you may later move to termly assessments. This balances collecting enough information to inform programme decisions while avoiding over-assessment, unnecessary costs and the challenge of detecting meaningful growth over very short periods.

- **Internal work schedule and coordination**: Create a comprehensive work plan with internal teams and external agencies covering all phases from enumerator training to data analysis and reporting. Ensure this plan explicitly details the logistical requirements and local stakeholder support (e.g. transport, venue access) required to carry out the assessments.

- **Prepare for pilot:** Start by selecting a small group of participants with clearly defined characteristics, so that the pilot offers a realistic preview of how the assessment will run in practice. Use the same tools, procedures and data collectors you intend to deploy at full scale. After the pilot, analyse the data to identify patterns, anomalies and any discrepancies between intended and actual outcomes. This will help you check both the rigour and feasibility of the tools and overall assessment design. Use these insights to refine your data collection approach and processes before full roll-out.

GLOBAL SCHOOLS FORUM

## Choosing learners to assess

Unless your programme has sufficient resources to assess every participating child, you will need to strategically select a representative sample of students for assessment. Selecting them in a systematic way ("sampling") helps ensure that your results reflect the wider group of students you support.

> **1. Select regions and schools**
>
> Unless your programme has sufficient resources to assess every participating child, you will need to strategically select a representative sample of students for assessment. Selecting them in a systematic way ("sampling") helps ensure that your results reflect the wider group of students you support.
>
> **2. Ensure key groups are included**
>
> Next, make sure the sample includes representative numbers of important groups of learners—especially students in different grades and genders. This is called **stratified sampling** and means selecting students separately within each group so that each group is represented in the sample and can be analysed. In practice, this is usually done by splitting the class register into the relevant groups (for example, boys and girls, or Grade 1 and Grade 2) and then selecting students within each group using the same method—such as choosing every xth student from the list. This ensures that the final sample reflects the different groups of learners in the population and allows you to compare outcomes across them.
>
> **3. Select new students for each round**
>
> Finally, repeat this sampling process each time the assessment is conducted. In many monitoring systems, assessments are **cross-sectional**, meaning a new group of students is selected for each round rather than following the same students over time. This avoids the logistical challenges of following the same children over time, while providing a representative snapshot of learning levels at each point in time.

The number of students you select (and how many schools they come from) matters because if you assess too few, you cannot be confident that the results reflect true learning levels across your programme. Sample size is always a balance between rigour and resource constraints, such as time, cost and the burden placed on schools and assessment teams. How large your sample will be that depends on the purpose of the assessment and the level of confidence you need in the results.

| Monitoring | Evaluation |
|---|---|
| For routine programme monitoring, there are no fixed rules for sample size. As a practical rule of thumb, assess at least 100 learners-- or all your learners, if you are supporting less than 100. If you are serving a larger number of children, aim to assess up to 10% of them, but no more than 1000. The number you choose to assess should, depend on how precise you want your estimate to be and available resources. | For programme evaluation with a comparison group, sample size depends on factors like how much you expect students to improve (effect size), how similar students are within the same school (intraclass correlation coefficient), and how confident you want to be in your findings (power and significance levels). |

General guidance is available in the EGRA Handbook, EGMA Handbook, and Save the Children's **Sample Size Calculator Comparison and Decision Making Tool**.

## Selecting mode of data collection

Digital assessments are excellent for 1:1 (individual) administration and can support adaptive or personalised testing, while paper-based assessments are more flexible for both 1:1 and 1:many (group) administrations, however, lack automation. Your choice of the data collection mode should be guided by factors such as:

- Assessment type (oral vs. written, 1:1 vs. group administration),
- Infrastructure availability (devices, power, connectivity),
- Data collection team capacity (digital skills, familiarity with tools),
- Budget (upfront and recurring costs).

EGRA, EGMA, and TaRL Assessments can be administered in either mode, so the decision should be made based on the implementation requirements outlined below. Self-administered EGRA must be paper based.

| Modes | Why choose this? | Implementation Requirements: |
|---|---|---|
| Digital (tablet/phone app) | Digital data collection reduces manual errors through built-in validation rules, dropdown menus, skip logic and automatic timers for fluency tasks. It also enables more personalised and adaptive workflows, for example, routing children to easier or harder items based on their responses or flagging specific learners or schools for follow-up. Integrated dashboards and exportable reports make it easier to analyse results rapidly and share stakeholder-tailored summaries. | • Functioning devices with sufficient storage, reliable charging options and plans for safe transport, maintenance and repair.<br>• Offline functionality where connectivity is weak<br>• Basic digital literacy among enumerators, targeted training.<br>• Cloud-based or institutional servers for secured data storage, automated backups and access controls.<br>• Pilot testing to verify platform stability, audio recording capabilities (for verification) and data encryption. |
| Paper-based | This is a more viable in remote locations with limited connectivity, lower up-front cost and easier roll-out with varied enumerator skill levels.<br>Forms are printed, manually administered and scored on-site, with subsequent digitisation for analysis. | • Clear data entry protocol (specifying who enters the data, where, by when)<br>• Rigorous quality controls, including random verification of entries against originals and automated checks to reduce errors<br>• Secured transport, storage of paper forms to prevent loss or tampering |

**Tangerine, Kobo Toolbox and Survey CTO** are some common digital data-collection platforms used for FLN assessments. Data can be collected offline and get synchronised when internet connection is available.

Irrespective of the mode, there is a need to standardise the approach by using standard versions of all tools and scoring guides, along with consistent respondent IDs that allow for seamless data matching across different collection points. The selected mode also needs to be piloted during tool validation to confirm operational feasibility, with results informing final data collection strategy.

GLOBAL SCHOOLS FORUM

**Case study: Data collection mode**

## Inspiring Teachers' SmartCoach app for student assessments and coaching cycles

In Ghana's Cape Coast Metropolis, Inspiring Teachers (IT) supports government school teachers to deliver the Inspiring Reading Program (IRP). The programme combines daily one-hour reading lessons with weekly formative assessments of taught skills. The SmartCoach app serves as a central tool to the teachers for carrying out termly 1:1 Oral Reading Assessments, which track child-level reading data and feed into a centralised dashboard. This provides Head Teachers (HTs) and School Improvement Support Officers (SISOs) real-time insights into lesson delivery quality and student progress.

The approach builds on Inspiring Teachers' earlier pilot of the model across 56 low-fee private schools, where an A/B test showed that giving HTs a coaching role supported by the SmartCoach App and training, led to improved teachers' delivery on structured pedagogy programme. An external RCT further confirmed accelerated learning in students, amounting to 0.5 SD gains on EGRA and reduction of non-readers by 44% within a year. Through the Labs, IT is now testing whether a similar tech and data-driven model can be successfully adopted and sustained by the district support system itself.

In low-fee private schools, Inspiring Teachers could provide direct oversight, while for government schools where teachers need more handholding, HTs have competing priorities and SISOs face logical constraints resulting in inconsistent school visits, the IT team had to build strong processes and support mechanisms to make SmartCoach app work.

In the government-led model, HTs conduct monthly classroom observations, while SISOs visit each school twice per term and conduct observations in SmartCoach. Teachers then receive feedback in the app, sent directly to them by the observer for reflections and improvement. Teachers continue to conduct ORF assessments through the app, reaching ~90% students on average, while coaching observations using the app is seeing a steady adoption.

To augment this, the IT team provides structured handbooks and comprehensive trainings to teachers, HTs and SISOs to leverage SmartCoach for both student assessments and classroom observations. Inspiring Teachers staff also conduct joint school visits with SISOs to provide field support and hold bi-weekly meetings with SISOs and district officers to review progress, identify schools needing support and course-correct. The team is also refining the app to prompt HTs and SISOs for action, schedule visits frequently and streamline coaching and assessment cycles - demonstrating that digitally collected student assessment data can inform both classroom instructions as well as wider district-level planning and supervision.

GLOBAL SCHOOLS FORUM

## Training data collection teams

Regardless of the data collection mode, data quality depends more on human execution, therefore, training is essential. Make sure you have in place a structured 4-5 days training for enumerators - for both new collectors and as a refresher for those who are skilled, so they are well capacitated and confident in undertaking the data collection process.

**Training agenda essentials:**

- **Orientation to ethical principles:** begin training with the core principles of consent, confidentiality and child safeguarding
- **Explain the "WHY":** show how data collection tasks links to the project's key indicators, final analysis and usage
- **Include mode-specific training support:**
  - **For Paper-based data collection:** include manual timing mastery (stopwatch use), form completion (legible written entries, standardised abbreviations), data entry and scoring consistency, skip pattern memorisation.
  - **For Digital data collection:** include device familiarisation and handling trainings (charging, offline sync, troubleshooting), app navigation and validation rules, steps to audio recording, photo upload, location tagging. Practice digital scoring and device workflow.
- **Conduct practical, hands-on exercises**- including live demonstrations and role-plays for trainers to model correct administration, practice in pairs and small groups exchange feedback. This can be further extended to practice in real settings in non-sampled schools or communities, followed by debriefs on common errors, corrections.
- **Ensure data consistency:** through Inter-rater reliability (IRR) checks. Have all enumerators assess the same child (or watch the same video) and compare scores to a "gold standard"; retrain anyone whose scoring differs beyond a pre-agreed tolerance.

## Good practices on data quality assurance

This example from Leadership for Equity (LFE) relates to classroom observation processes rather than student learning outcomes; however, the same quality assurance principles, i.e, clear protocols, regular checks, and verification of scoring consistenc apply when collecting learning assessment data.

**Leadership for Equity (LFE's) middle managers' training on the World Bank Teach tool in Maharashtra-India:**
To maintain data integrity while conducting classroom teaching observations by government supported observers, LfE adopted practices like:

- **Routine debriefs** between the Central M&E team and field teams (responsible for data collection), held daily or weekly throughout the data collection period, to address emerging challenges, clarify scoring doubts, and check submission completeness.

- **Learning logs** recorded monthly by the teams to capture any protocol deviations and contextual tweaks made during the testing process. Refer to a learning log tool for documenting ongoing insights and contextual tweaks made during the testing.

- **Methods like double-coding protocols were used,** where LfE team accompanied local education officials (Kendra Pramukhs) to observe the same classroom simultaneously. Each scored the observation independently across the Teach tool's nine elements. Following this, the scores were compared and discussions were held to reconcile differences where ratings diverged. Variances (typically ranging from 1-10%) between LfE team scores and those of the officials were transparently reported. Data points were deemed unreliable and excluded if more than three of the nine Teach tool elements showed a deviation exceeding ±1 point.

# 4. Analyse assessment results appropriately

Once data has been collected, it must be digitised (if not already captured electronically), securely stored, and cleaned before analysis begins. This includes checking for missing or inconsistent responses and documenting any corrections or exclusions, so that the process is transparent and can be reviewed for quality assurance. Those reviewing the data should be familiar enough with the assessment to recognise potential errors.

**How learning outcomes are analysed is just as important as what is measured.** Reviewing raw scores—such as gains in oral reading fluency—is one valuable way to understand progress. However, additional layers of analysis can help interpret these results more meaningfully. Different questions about learning require different types of analysis, so the type of analysis you choose should directly reflect your purpose for measuring foundational learning outcomes. For example, programme teams may want to know whether students met a minimum proficiency benchmark, whether students moved up one or more skill levels, whether the proportion of non-readers has decreased, or how students participating in their programme compare to similar students who did not participate in it. Each type of indicator has its own uses, strengths, and limitations, so choosing the right analysis ensures that the data you collect leads to useful and actionable conclusions.

## Benchmarks

Benchmarks are minimum standards that represent critical thresholds in the development of reading or mathematics. They indicate whether learners have mastered foundational skills sufficiently to support further learning and later success. Benchmarks are not universal; they are specific to the assessment, subject, grade, language, and educational context. Once a benchmark is defined, programmes can measure the percentage of learners who are at or above the benchmark (on track) and those who fall below it (not on track) and examine how this distribution changes over time. Because benchmarks represent critical thresholds, indicators based on benchmark attainment are especially important for assessing whether a programme is effective enough to "move the needle" and produce meaningful change. These are general guidelines for selecting benchmarks across the three most common FLN assessments.

- **EGRA**: EGRA consists of several separate tasks, many timed, often including identifying letter sounds or names; reading real or invented words; and reading a passage and answering comprehension questions about it. There is not a universal way to create a single composite score from these tasks, but of them, reading a passage is the best representation of overall reading skill. It is therefore typically used as the primary metric for benchmarking, while scores on other tasks provide supporting evidence of skill development. For the EGRA passage reading, learners read a grade-level text for one minute, and the enumerator counts the number of words they read correctly during that minute. This is called oral reading fluency (ORF). There are no universal benchmarks for EGRA because languages have unique scripts, word lengths, and sound-symbol relationships that impact reading rate. Instead, benchmarks are developed for specific languages and contexts, typically by finding the reading rate associated with correctly answering 4/5 comprehension questions about the text (although, see Ardington et al., 2021). Benchmarks for specific languages and contexts can often be found in national curricula, programme evaluation reports, and academic articles. For example, these are the ORF benchmarks set in Kenya's national curriculum:

## ORF benchmarks in Kenya's national curriculum (words correct per minute)

| **Language** | **Grade 1** | Grade 2 | Grade 3 |
|---|---|---|---|
| English | 30 | 65 | 90 |
| Swahili | 30 | 45 | 55 |

- **EGMA**: EGMA also consists of multiple separate tasks, typically including identifying numbers, identifying the larger of two numbers, identifying missing numbers on a number line, basic addition and subtraction problems, and word problems. Because mathematics curricula vary across countries, both in the pace at which content is introduced and in the order in which specific skills are taught, benchmarks are highly context-dependent. Benchmarks for specific subtasks can be established by mapping EGMA items onto national curricula, which is what GSF and Duara Education did to develop mathematics benchmarks for Duara's lower primary grade students in Kenya. In Kenya's national curriculum, students are expected to recognise and understand the value of numbers up to 50 by Grade 1, up to 100 by Grade 2, and up to 1,000 by Grade 3. Using that information, it is possible to determine the grade level of each item in the number discrimination task of the core EGRA Instrument. Out of the 10 items, first graders should correctly answer three, second graders six, and third graders all ten.

# Illustrative: Core EGRA Number Discrimination Sub-task

Instructions: Tell me which number is bigger.

| Grade Level Expectation | Core EGMA Item | Expected Number Discrimination accuracy for grade level |
|---|---|---|
| Grade 1 | 1) 7 or 5<br>2) 11 or 24<br>3) 39 or 23 | 30% |
| Grade 2 | 4) 58 or 49<br>5) 65 or 67<br>6) 94 or 78 | 60% |
| Grade 3 | 7) 146 or 153<br>8) 285 or 534<br>9) 623 or 632<br>10) 867 or 965 | 100% |

To create a composite EGMA benchmark, this curriculum mapping is carried out across all untimed EGMA subtasks. For each grade:

- Identify which items are expected to be mastered based on the curriculum
- Sum the total number of expected items across subtasks
- Divide by the total number of items to obtain a grade-specific benchmark percentage

Because occasional mistakes can be expected even when students are proficient in a skill, this benchmark percentage can be adjusted—for example, multiplied by 80%—to allow for a small number of errors. In the example above, this would mean that a third-grade student could answer eight of ten items correctly and still be considered at benchmark. Organisations interested in using EGMA benchmarks could apply a similar approach by mapping EGMA items to the national curriculum in their context and constructing grade-specific composite benchmarks that reflect local expectations.

- **TaRL Assessments**: General benchmarks exist for the standard TaRL Assessment format, but these may need to be adapted depending on the specific form used and the national curriculum expectations for each grade. Differences in the literacy and mathematics tasks included in an assessment can affect what constitutes an appropriate benchmark. The table below is illustrative only; programme teams should confirm benchmarks by consulting assessment developers and reviewing national curricula.

GLOBAL SCHOOLS FORUM

## Standard TaRL Assessment benchmarks by grade level

| Subject | Grade 1 | Grade 2 | Grades 3+ |
|---|---|---|---|
| Reading | Paragraph | Story | Story |
| Mathematics | Number recognition | Subtraction | Division |

✓ Benchmarks show how many learners have reached a defined level of proficiency—that is, how many have a strong enough foundation to be on track for later success.

✗ In contexts where overall performance is very low, benchmark analysis alone can miss important progress among learners who are improving but have not yet reached the benchmark.

## Levels

Level-based indicators show the percentage of learners at each performance level (for example, non-readers, letter readers, word readers, or paragraph readers). These indicators can be used to track how the distribution of learners across levels changes over time—for example, an effective reading programme might see a reduction in the share of learners reading letters only, and an increase in the share reading words or paragraphs. This helps programmes assess whether learning improvements are meaningful across the full range of learners. Level-based indicators support equity analysis by showing who is benefiting from the programme - whether gains are concentrated among the most disadvantaged learners or primarily among those who started higher. When learner-level data are available, level indicators can also be used to identify how many individual children made meaningful progress.

✓ Level-based indicators show how learners are distributed across skill bands, making it possible to see growth among both lower- and higher-performing learners.

✗ Levels are only as useful as the cut off points used to define them. When thresholds are not meaningful, small score changes can shift learners between levels without much real change in skills.

## Zero reduction

Zero reduction indicators measure the share of learners who cannot complete any tasks on the assessment and can be used to track whether that group is shrinking over time. It is most commonly used for ORF. Zero reduction indicators can help show whether the lowest-performing learners are beginning to acquire foundational skills. This type of indicator is often used when overall learning levels are very low and one of the programme goals is to ensure that no learners are completely excluded from gaining foundational skills.

Zero reduction is easy to understand, comparable across contexts, and shows the impact of programmes on the lowest-performing learners.

It captures only movement out of the very lowest category, does not reflect progress among learners who start at higher levels, and does not indicate how many students are reaching or nearing proficiency.

**Case study: Analysis design**

**Beyond averages- what multiple indicators revealed in Room to Read's (RtR) pilot programme**

Room to Read-Tanzania implemented a remedial literacy pilot programme in Dar es Salaam, in two distinct settings – i) children's homes- where students receive after-school reading support within a residential setting, and ii) MEMKWA-COBET centers- providing alternative education for out-of-school children and youth. Both these settings serve children who have experienced educational disruption or limited academic support, where most of them begin with very little or no foundational literacy skills.

Since children's starting points varied considerably, a single measure metric would have obscured as much as it revealed. To ensure methodological rigor, the program's impact was evaluated across 669 student assessments (Baseline N=337; Endline N=332), tracking progress across 337 unique learners. The programme therefore tracked multiple indicators to understand how the pilot programme affected different groups of children and different components of reading development across the learning distributions.

- **Zero-Score Reduction** (Capturing movement at the bottom): This metric tracks the percentage of students unable to score a single point on a given subtask. For example, in oral reading fluency, 25% of Children's Home learners could not read a single word at baseline, dropping to 10% by endline. Across all subtasks and settings, the programme achieved an average 18.5 percentage point reduction in zero-scores, headlined by a 37-percentage point reduction in letter-sounding zero-scores. This movement proves the programme effectively helped the most marginalised learners exit the "non-reader" category and begin developing foundational decoding skills.

- **Tracking Cohort Momentum** (The Middle): Beyond binary categories, the programme measured mean score gains across all six subtasks to capture incremental progress for the entire group. Overall, Oral Reading Fluency (ORF) rose from 18.27 to 28.48 correct word per, minute (cwpm) while Sentence Dictation scores more than doubled, increasing from 3.0 to 6.3 (out of 12). This shift indicates that even learners who had not yet reached the benchmark were making significant, measurable gains in foundational literacy.

- **Benchmark Attainment** (Capturing movement at a defined proficiency level): This indicator tracks the percentage of learners reaching a rigorous reading benchmark (defined as 45+ cwpm and 80%+ comprehension). Student progress was measured using the Early Grade Literacy Skills Assessment (EGLS-A), a six-task tool developed and localised for Kiswahili, aligned to the Tanzanian national curriculum, and validated through two rounds of large-scale pilot testing by RtR. Across both settings, benchmark attainment for oral reading fluency (45+ cwpm) more than doubled overall, increasing from 8% to 21%. In Children's Homes specifically, the proportion of learners reaching the benchmark more than tripled, increasing from 6% to 20%, while MEMKWA-COBET centers saw an increase from 10% to 22%. This upward trajectory suggests the gains were not confined only to the struggling readers but supported a significant portion of the cohort in reaching a critical level of reading proficiency necessary for continued learning.

Together, these three tiers of measurement enabled RtR to track diverse types of progress across the entire learning distribution.

## Mean comparisons

Mean comparisons look at average scores for the groups of learners participating in a programme and compare them to the scores of learners who did not participate. This type of analysis shows whether learners in the programme improved more, on average, than learners who did not receive the programme. Mean comparisons allow you to calculate an effect size, which is a standardised measure of how big the difference is between groups. It is typically reported in standard deviations.

Effect sizes are useful because they allow learning gains to be compared across programmes and contexts. However, effect sizes are influenced by factors unrelated to programme quality—such as learners' starting levels and the spread of learners' skills—which can make gains appear larger or smaller than students' real progress. For this reason, effect sizes should be interpreted alongside other indicators, especially benchmark attainment, which is more representative of whether learners have reached levels of proficiency that matter for schooling and further learning.

✓ Mean scores provide a simple summary of overall performance and allow for clear comparisons across groups or over time.

✗ Averages can mask differences within the group, hiding whether gains are concentrated among a few learners or shared by all.

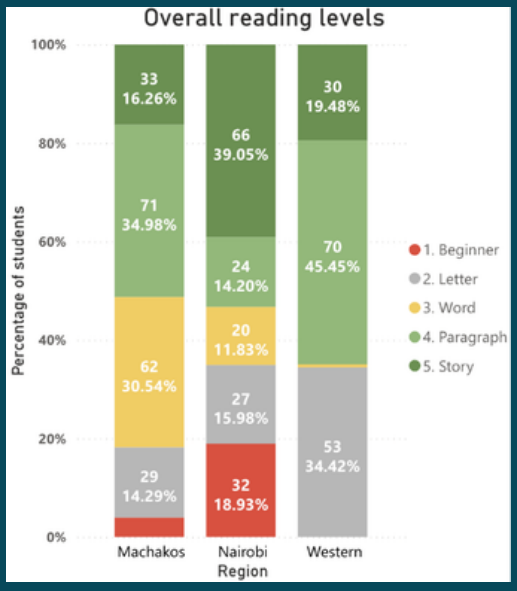
The table below illustrates several common ways to analyse foundational learning data and the kinds of questions each approach can help answer. The example indicators, proficiency levels, and results are drawn from the PRIMR and Tusome programmes in Kenya. All four types of analysis can be applied to EGRA, EGMA, and TaRL Assessments; the table shows selected examples only.

| Type | When to consider using | Indicator | Proficiency | Result |
|---|---|---|---|---|
| Benchmark<br>Using EGRA | You want to know what percentage of students are on track for future success | % of learners who read a Grade 2 level text in Swahili at an appropriate rate | Reading 45 words correct per minute | 4% in 2015<br>12% in 2016 |
| Levels<br>Using TaRL Assessment | You want to know how many students improved meaningfully | % of learners with improved reading skills following participation in remedial reading programme | Proficient with letters, words, paragraphs, or stories | 67% of students improved by at least 1 level |
| Zero reduction<br>Using EGRA | You want to know how the programme impacted students with the lowest baseline scores | % of Grade 1 learners unable to read at least one word of a passage | Reading at least one word | Zero scores decreased from 53% to 23% |
| Mean comparison<br>Using EGRA | You have a comparison group and want to compare improvement between the two groups | ORF Word Count Per Minute (WCPM) for Grade 1 English in non-formal schools | N/A | The group receiving the programme improved by 19 words correct per minute more than the comparison group (effect size = .51) |

Another critical part of analysing data is examining at differences in outcomes across key groups of learners to understand whether a programme is benefiting students equitably. This typically involves disaggregating results by important characteristics for the context, which might include age, gender, geography (for example, rural vs. urban, or different regions), language background, socioeconomic status, school type, or disability status. Looking at results in this way helps identify whether certain groups are benefiting less from the programme or facing additional barriers to learning. The analysis is usually conducted by calculating the same indicators (for example, zero scores or proportion at benchmark) separately for each group and comparing the results. If gaps between groups appear, they can point to areas where the programme may need to adapt implementation, provide additional support, or investigate contextual factors affecting participation and learning.

## Case study: Disaggregated Analyses

One example of this kind of analysis comes from **Angaza Elimu, in Kenya, who examined differences in the distribution of reading skill levels across the three regions where they operate:** Machakos, Nairobi, and Western Kenya. This comparison revealed differences in learning profiles across locations. In Nairobi, an urban area, a larger share of learners was concentrated at the two extremes: many students were at the very lowest level (unable to recognise letters), while a substantial proportion were already able to read stories. In contrast, in the more rural regions of Machakos and Western Kenya, a larger share of learners fell in the middle levels of reading development, like being able to read words or paragraphs.




Comparison of student reading levels across three regions (Angaza Elimu Baseline Data)

These differences suggested that learners were entering the programme with distinct learning profiles depending on the context. As a result, Angaza Elimu began investigating possible factors that might explain these patterns. This included examining similarities and differences in teacher experience, classroom conditions, and learner characteristics across regions, as well as language backgrounds and the languages used for instruction.

Understanding these contextual differences helped the organisation tailor coaching and instructional support to the needs of teachers and learners in each region.

# 5. Connect data to decision-making

Assessment data alone does not change outcomes. What matters is how data and insights from assessments are presented, discussed, and acted upon within organisations through structured feedback loops. When strategically applied, assessment data can be used to continuously refine programmes – whether to adjust curriculum content, modify teacher training plans, redesign tools or modify implementation strategies.

**Instead of treating data as an endpoint for reporting, these insights should be used for adaptive decision-making, guiding classrooms practices, policies and programme frameworks.**

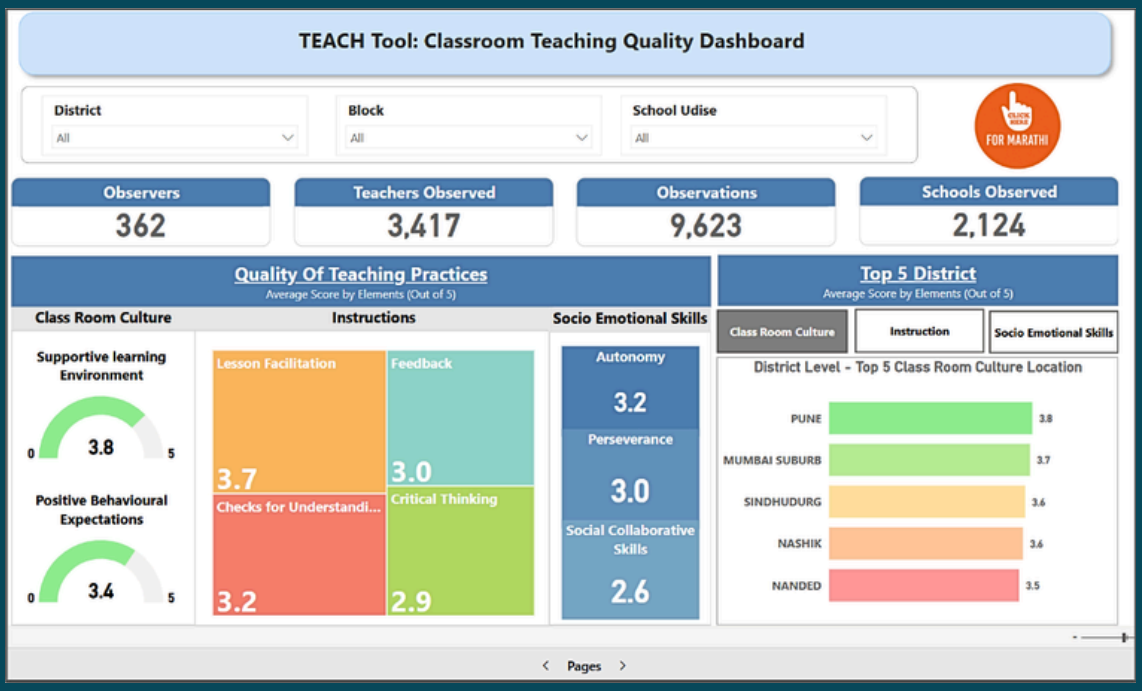
**Case study: Feedback loops**

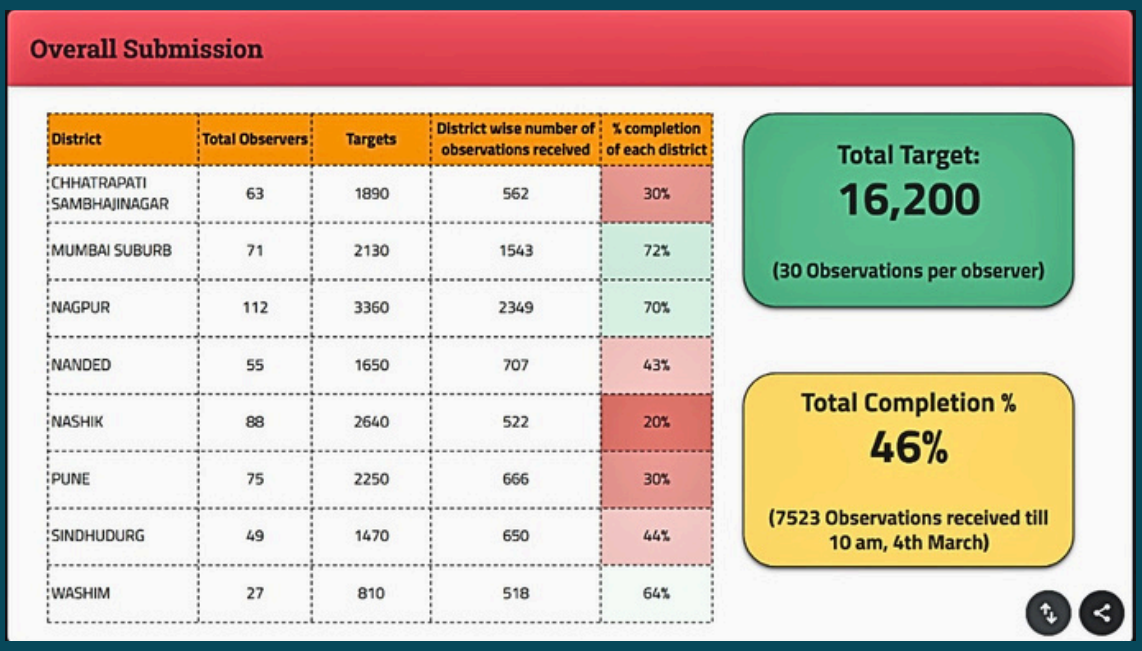
### From student assessments to system-wide action: Leadership for Equity's (India) data-backed decision-making approach

Turning classroom data into system-wide action requires the right data, reaching the right people, at the right time. Leadership for Equity's (LfE) structured cycle starts from cluster and block officials (Observers) collecting student learning and teacher practice data from classrooms through a digital form on their phones. Three types of data are collected: (a) student learning assessments (including both, baseline and formative assessments), (b) classroom observation data (teacher practice, lesson delivery, student engagement), and (c) observer activity (who observed, when, how many observations, and observer ratings).

The performance against key metrics (see table below) is visualised through new or existing dashboards at three levels: block, district and state. A Block Education Officer, for instance, sees an aggregated view of all observations across their block including key trends, outliers and which observers haven't submitted. The design of the dashboards must consider 'What decision does this level need to make?' rather than 'What data do we have?'

## Case study: Feedback loops




| District | Total Observers | Targets | District wise number of observations received | % completion of each district |
|---|---|---|---|---|
| CHHATRAPATI SAMBHAJINAGAR | 63 | 1890 | 562 | 30% |
| MUMBAI SUBURB | 71 | 2130 | 1543 | 72% |
| NAGPUR | 112 | 3360 | 2349 | 70% |
| NANDED | 55 | 1650 | 707 | 43% |
| NASHIK | 88 | 2640 | 522 | 20% |
| PUNE | 75 | 2250 | 666 | 30% |
| SINDHUDURG | 49 | 1470 | 650 | 44% |
| WASHIM | 27 | 810 | 518 | 64% |

Dashboard developed by LfE with government systems, discussed during review meetings

**Case study: Feedback loops**

# Key Performance Indicators tracked across different levels

| Focus Areas | Leading Indicators | Frequency | Owner |
|---|---|---|---|
| Student Learning | Assessment scores by schools/block/district | Monthly | District and Block Education Officials |
| Teacher Practice | Observation ratings per teacher | Per cycle/Term | Cluster Officials |
| Cluster/Block Official (Observer) Engagement | No. of observations completed vs. target | Monthly | Block Education Officials |
| Data Quality | Inflation flags, consistency checks | Per cycle/term | Local Officials and Field Team |
| Review Meeting Adherence | BQCS/DQCS held on schedule | Monthly | District Admin |
| Action Follow-through | % action points closed from last meeting | Monthly | District and Block Education Officials |

LfE embeds data discussions within existing monthly Block and District Quality Cell meetings (BQCS/DQCS). The District Education Officials chair the review discussion which is centred around reviewing performance of schools at a block level based on SLO scores, identifying schools that need support, and selecting the ones who deserve recognition. Top performers are celebrated in front of peers, creating positive social pressure as well as course-corrections made on areas of improvements.

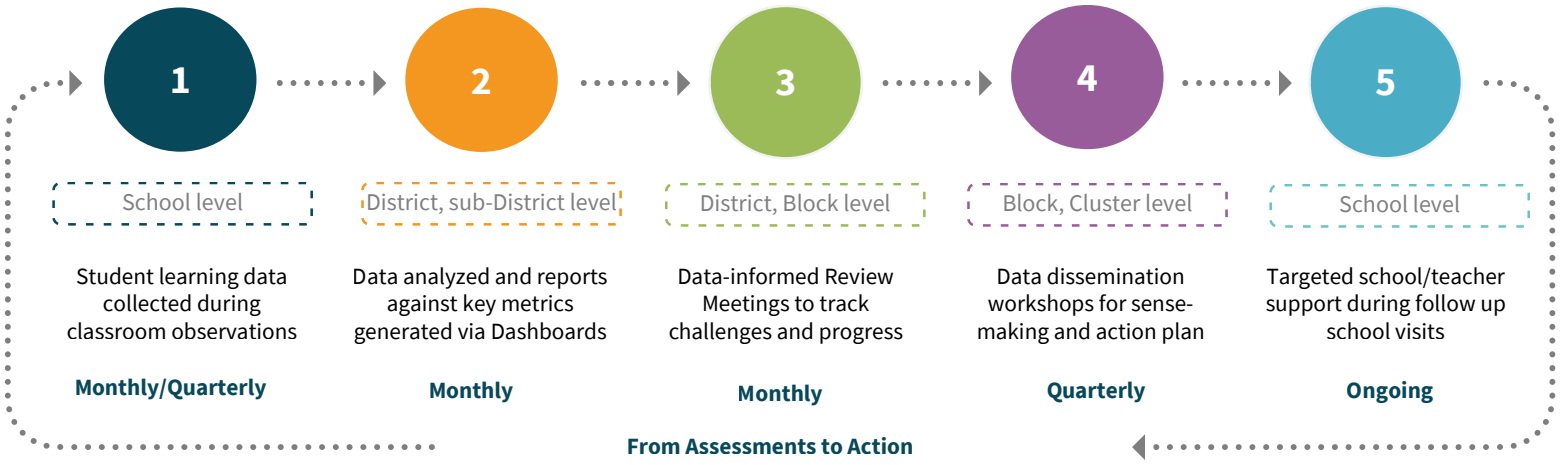

Sense-making workshop underway to deep-dive into data for action | Image Courtesy – Leadership for Equity

At the end of each programme cycle, LfE convenes data dissemination workshops to deep-dive on data with district officials. This feeds directly into action planning wherein, block and district officials identify which schools and teachers need support, assign responsibilities and set visit targets for the next cycle. The next cycle's data is then compared against the last focussing on observer compliance, data quality and learning outcomes. This enables continual data circularity embedded within system's exisiting structures and resources.

GLOBAL SCHOOLS FORUM

# Classroom to System Feedback Loop

Data-backed decision making is key to improve student learning at scale. By coupling strong data infrastructure with deliberate human processes scattered numbers can be put into understanding and understanding into action - for continuous improvement.




## Key considerations while developing structured data-feedback loops

**Ground data in reality, with the right people in the room**. Sharing student-level data regularly helps different levels of the system "see the classroom" and align decisions closer to learning realities. Learning outcome data becomes far more actionable when viewed alongside implementation data, such as classroom observations, coaching visit frequency, attendance records and other fidelity measures that help explain why outcomes are or aren't improving. Crucially, involving people with real decision-making authority from the outset means they help shape the questions and indicators being tracked, are more likely to trust the findings and understand what actions are feasible within existing constraints.

**Design for the decision, not the dataset:** At every level, ask: what decision does this stakeholder need to make and what is the minimum data they need to make it well? This prevents overload and ensures clarity over noise. It is equally important for timely insights to reach implementers while there is still time for course correction. Data sharing should be timed to key periods: when academic sessions are starting or are at peak, budgets are still available or training cycles are being planned. This is especially important for foundational learning, where early delays compound quickly.

**Embed data review into system's existing rhythms:** Integrating data discussions into review meetings that are already scheduled at national, district and sub-district levels ensures regularity and signals that data is core to leadership and not a burden. Similarly, leveraging existing system dashboards with data access based on the scope of decision making and support- at the state, district, block levels help, so that everyone in the system is looking at the same data driving decisions in real time.

- **Build an improvement culture:** Review meetings should establish clear expectations upfront such as when and how data will be reviewed, what will be shared and how findings will be interpreted collectively. Early sessions should focus on surfacing biases and building trust. Meetings that only surface problems incentivise data manipulation; consistent recognition of high performers shifts the culture toward one where good data leads to good outcomes.

- **Secure commitment to action:** Effective feedback loops close only when data leads to explicit decisions about what each actor will do differently towards improving student learning - based on the evidence. For example: Can government coaches adjust their classroom observation frequency in underperforming schools? Can teachers build targeted teaching-learning plans from formative assessment data? Can implementing organisations reallocate resources differently across levels of the system to bring greater clarity and accountability? Without this specificity, data risks remaining descriptive rather than transformative.

## Conclusion

The evidence base for FLN assessment has matured considerably over the past two decades. We have validated tools in many languages and contexts, developed guidelines for interpreting the results of common assessment formats, and increased focus on measuring outcomes rather than just inputs. And yet, in many FLN programmes, measurement is under-resourced or disconnected from programme decisions; as a result, programmes miss opportunities to learn from evidence and improve their effectiveness in closing learning gaps for children.

To address this, the guide provides actionable guidance tailored to programme realities. It frames measurement as a core part of implementing programmes, and provides guidance on what to assess, how to collect and analyse data, and how to translate that data into practical decisions that can directly shape children's learning trajectories and ensure no learner is overlooked.

When measurement is embedded in programme cycles, it helps organisations understand whether their approach is working, identify which learners need additional support, and adjust implementation accordingly. Strengthening how we measure and how we act on what we learn is, therefore, a critical step toward ensuring that FLN programmes deliver meaningful learning gains for the children they serve.

# Appendix I: Validity and Reliability

Even when an assessment meets practical needs—such as being simple to administer and low-cost—it still must produce results that accurately reflect what students know and can do. Two core psychometric concepts help determine whether data from an assessment can be trusted: validity and reliability. **Validity** indicates whether the test measures the skill it is intended to measure, while **reliability** refers to whether the results are consistent across students, time, and test forms. The table below outlines the key types of validity and reliability that are most relevant when selecting assessments for foundational literacy and numeracy.

## Validity

| Term | Question to ask | Why it matters | FLN example |
|---|---|---|---|
| **Face validity** | At first glance, does the assessment seem to measure the skill it claims to measure? | When an assessment looks appropriate and relevant, users are more confident that it measures what it claims to measure. | If a literacy assessment for first graders consists mostly of reading complex passages, it lacks face validity because it doesn't seem appropriate for their level. |
| **Content validity** | Does the assessment cover all the important content areas needed to measure this skill? | Covering all key content areas ensures the assessment reflects the full scope of what learners should know or be able to do. | An early math test that does not assess number identification would lack content validity, as it is missing a key area. |
| **Construct validity** | Does the assessment truly measure skill it is intended to assess, rather than something else? | This ensures the test isn't accidentally measuring something else, so your conclusions about students' abilities are accurate. | A written math test whose scores are very highly correlated with a reading test may lack construct validity—it may be measuring reading skills rather than math skills. |
| **Criterion validity** | How well do the assessment results align with another trusted measure of the same skill? | When the results align with the results of a trusted assessment, you can have similar confidence in the results. | A new reading assessment that is highly correlated with the Early Grade Reading Assessment, a previously validated "gold standard" assessment, demonstrates strong criterion validity |

GLOBAL SCHOOLS FORUM

## Reliability

| Term | Question to ask | Why it matters | FLN example |
|---|---|---|---|
| **Test-retest reliability** | How consistent are results when the same students take the assessment at two different times? | Consistent scores over time show the results are stable and not overly affected by unrelated issues like time of day, mood, or distractions. | If a group of students take the same reading assessment two weeks apart and their scores are very similar both times (assuming their skills have not changed much), the assessment shows strong test–retest reliability. |
| **Inter-rater reliability** | How consistent are results when different people administer or score the assessment? | When different assessors give similar scores, you can trust that the results reflect learners' skills rather than individual assessor differences. | If two different assessors listen to the same student read aloud and both record nearly the same number of correct words per minute, the assessment demonstrates strong inter-rater reliability. |
| **Parallel forms reliability** | How consistent are results when different versions of the test are administered? | High parallel forms reliability shows that no single version of the test is easier or harder, making results across forms comparable. | If two different versions of a mathematics assessment contain different items and students score about the same on both versions, the two versions show strong parallel forms reliability. |
| **Internal consistency** | How strongly are scores on different test items related, showing they measure the same skill? | High internal consistency shows the items are measuring the same skill, so the overall score is an accurate indicator of that skill. | If students who correctly answer one additional problem also tend to answer other additional problem correctly, the assessment shows strong internal consistency because the items measure the same skill. |

Your experience with this Practitioners' Guide matters and we would love to hear from you. Whether you have feedback, questions or want support in applying it to your context, we welcome the conversation:


**sarah.rotich@globalschoolsforum.org**


**ratul.chowdhury@globalschoolsforum.org**

GLOBAL SCHOOLS FORUM

## Become a member

Join our growing community of innovative organisations dedicated to transforming education

## Visit our website

www.globalschoolsforum.org

## Email us

info@globalschoolsforum.org

## Follow us