What Works Hub
for Global Education

Measuring implementation in education

Insight Note

Noam Angrist, Sharnic Djaker, Sam Ho, Michelle Kaffenberger, Guillermo Romero, Kate Sturla

March 2026

BLAVATNIK SCHOOL OF GOVERNMENT
UNIVERSITY OF OXFORD
UK International Development
Partnership | Progress | Prosperity

Measuring implementation in education

Insight Note

Noam Angrist
noam.angrist@bsg.ox.ac.uk

Sharnic Djaker
sharnic.djaker@bsg.ox.ac.uk

Sam Ho
sam.ho@bsg.ox.ac.uk

Michelle Kaffenberger
michelle.kaffenberger@bsg.ox.ac.uk

Guillermo Romero
guillermo.romero@bsg.ox.ac.uk

Kate Sturla
kate.sturla@bsg.ox.ac.uk

The What Works Hub for Global Education is an international partnership, funded by the UK government's Foreign, Commonwealth & Development Office and the Gates Foundation, working out how to effectively implement education reforms at scale.

Please cite this as:

Angrist, N., Djaker, S., Ho, S., Kaffenberger, M., Romero, G. & Sturla, K. 2026. Measuring implementation in education. What Works Hub for Global Education. Insight note. RI_2026/005. **https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-RI_2026/005**

Table of contents

Introduction

Whether an education programme succeeds or fails often hinges on the quality of implementation. Did teachers show up to the training? Did the school support officers offer the type of mentorship that was planned? Did new textbooks reach schools and get used in the classroom? These factors are critical to whether children achieve intended learning gains. Yet implementation is often understudied and underreported: Fewer than 12% of impact evaluations in education measure and account for implementation (Angrist & Meager, 2023). Moreover, there is little consensus on how to measure implementation and which dimensions to prioritise on a limited evaluation budget (Ryan et al., 2024).

Implementation measurement is crucial to help us (a) diagnose why some interventions fail, (b) understand why an intervention may succeed in some settings more than others, and (c) course-correct implementation challenges in real-time.

Recent studies demonstrate the returns to systematic approaches to quantifying and accounting for implementation. For example, Angrist & Meager (2023) show that variation in implementation rates is a central factor in explaining heterogeneity in treatment effects for teaching at the right level programmes – which vary by an order of magnitude across contexts – with implementation explaining nearly all the variation, mattering more than other factors such baseline conditions or geographic context. D'Agostino et al., (2024) similarly find that differences in implementation such as frequency of coaching visits explain much of the variation in programme effects across school sites. Angrist & Dercon (2024) compare policy plans with household reports on the receipt of services, identifying large implementation gaps. These studies, among others, underscore the importance of quantifying implementation in education.

In this guidance note, we put forward a practical framework for measuring and reporting implementation in education. We also share a toolkit and resource package the What Works Hub for Global Education is developing to support implementation measurement. The aim with this framework is to support a more systematic approach to implementation measurement to help unpack the 'black box' of how educational interventions achieve impact at scale, and to ultimately advance learning outcomes for all children.

Several related efforts emphasise the importance of monitoring implementation including the *Goldilocks Toolkit: Monitoring for Learning and Accountability* by Innovations for Poverty Action (IPA) (2016); the U.S. Department of Education guide for researchers, *Conducting Implementation Research in Impact Studies of Education Interventions* (Hill et al., 2023); and the OECD's *Implementation Framework for Effective Change in Schools* (2020).

Despite this growing interest, consistent measurement of implementation within education has remained limited. A recent scoping review of 314 papers by Ryan et al., (2024) found little consistency in implementation terminology or framework adoption. Moreover, current efforts focus on a limited set of implementation concepts, often at higher policy levels. We build on this work to provide a more comprehensive framework, including a focus on the front line of implementation (e.g., the classroom) and a set of concrete and aligned measurement tools.

1. A framework for measuring implementation

1. Implementation concepts

1.1 Implementation stages, agents, and units

Education takes place within layered systems that connect policy to practice. Decisions and activities often flow through several levels of an education system – from national policymakers to district administrators, school leaders, teachers, and finally students. We define several concepts below to make implementation concrete.

- ***Implementation stages.*** We call each step in the implementation chain a *stage*. These stages are sequential steps through which an intervention flows from policy down to delivery to students, typically in schools and classrooms.
- ***Upstream and downstream stages***. We consider earlier stages of policy administration as occurring *upstream*, while the classroom and student experience represent *downstream* stages of implementation.
- ***Agents***. In each implementation stage, there is an implementing *agent* who is responsible for delivery. These may include ministry officials, district officers, headteachers, and teachers, among others. In many cases, the recipient in one stage becomes the delivery agent in the next (for example, a teacher who receives training in a new pedagogy then delivers it in the classroom). Ultimately, the final agent is the student who benefits from an intervention.
- ***Units***. At each implementation stage, there is a well-defined unit where delivery occurs. Units represent the organisational or physical levels at which implementation takes place, such as districts, schools, classrooms, or households.

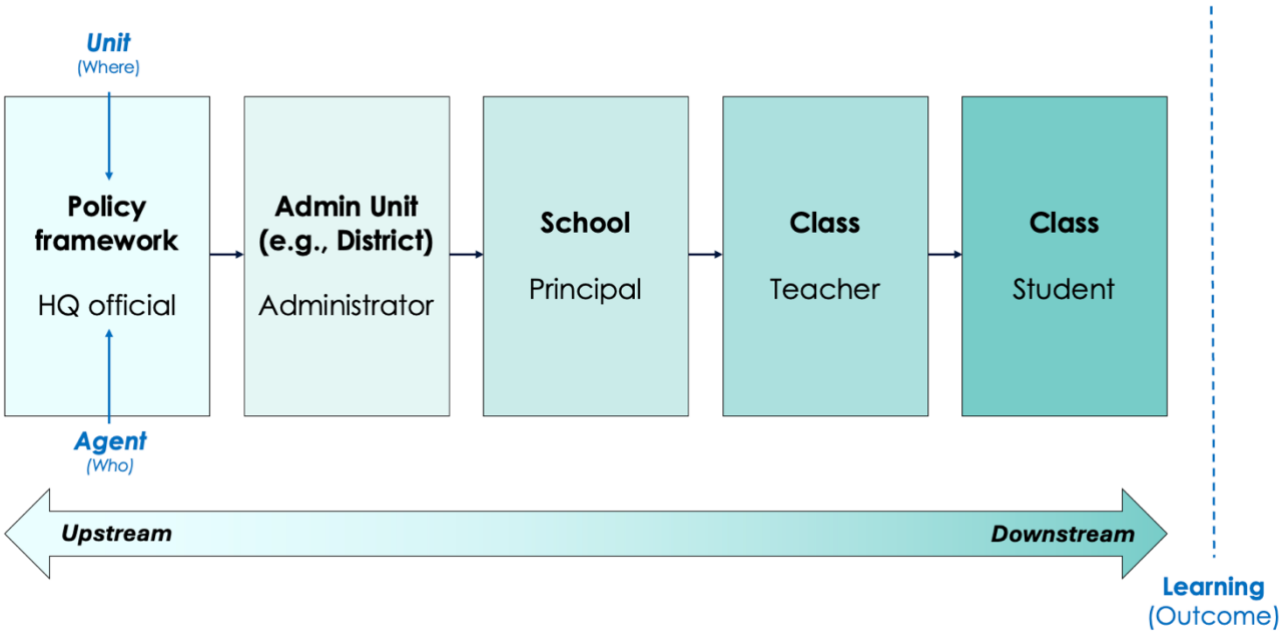


Figure 1: Implementation stages, agents, and units in an education system

Figure 1 provides an illustrative diagram of implementation stages, agents, and units. This diagram traces how an intervention and its components flow through a system – i.e., where it begins, how it transitions between stages in the system, and where it ultimately concludes.

Of note, not all interventions are delivered within the school system. Some interventions operate outside the school and are delivered through a broader education ecosystem, including the household (unit) and the parent (agent). In such cases, the same principles apply; one can delineate the stages, agents, and units involved in the implementation delivery system.

***Box 1:** A note on flow*

> While we depict education *interventions* as flowing from upstream to downstream, *data* can often flow downstream to upstream (see Figure 2). Successful monitoring should result in positive feedback loops (i.e., data flowing upstream to improve implementation flowing downstream).

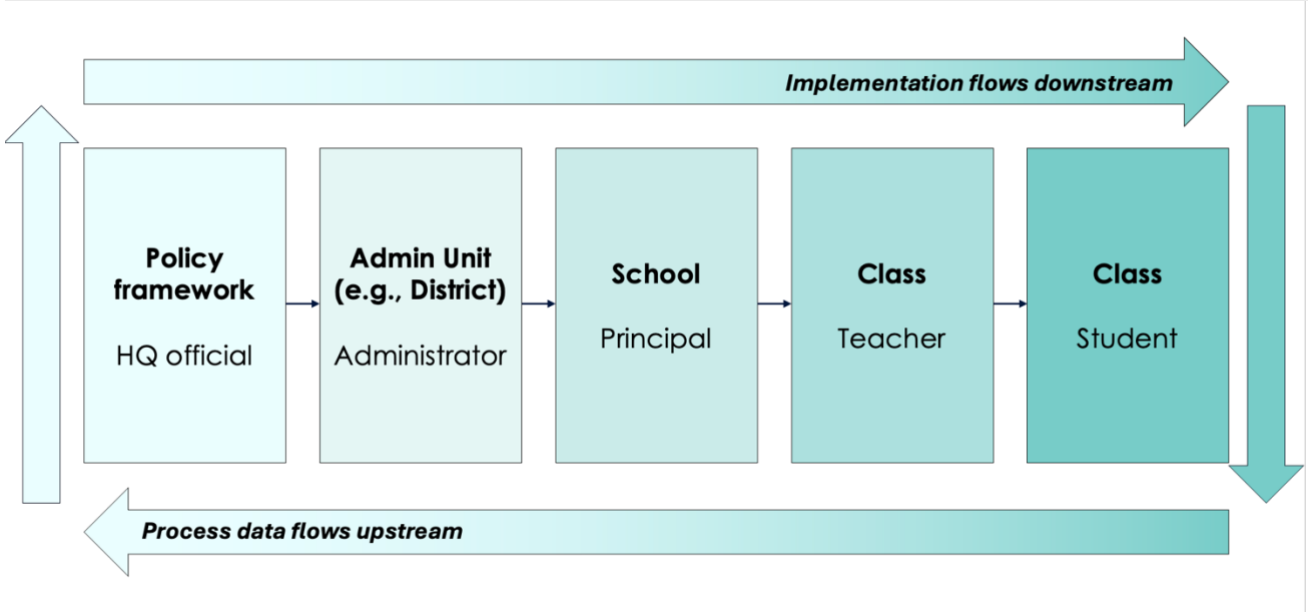


Figure 2: Interventions often flow downstream, data often flows upstream

1.2 Programme components

Programme components are the essential activities that form the intervention. These are often similar to the elements of a theory of change. It is critical to define these components upfront in order to make the intervention clear and concrete. The design of education interventions often consists of several programme components (see Figure 3). For measuring implementation, programme designers will ideally articulate five to ten main components of an intervention throughout the entire implementation chain. To keep measurement focused and parsimonious, it is important to prioritise by focusing on major components, rather than creating an exhaustive list of smaller programme features. For example, major components of a targeted instruction intervention might include teacher training and ongoing coaching at the school level (upstream); and regular assessment and grouping by level at the classroom level (downstream).

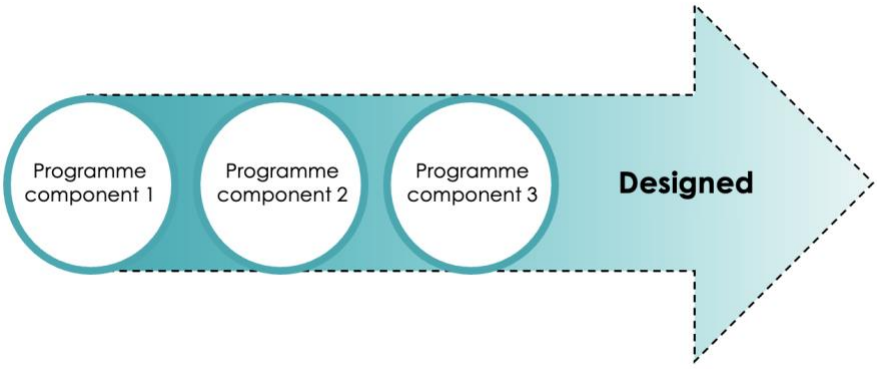


Figure 3: An intervention design is comprised of several programme components

Once defined, programme components can be mapped to their respective stages in the implementation measurement framework. Below we include an example from a teaching at the right level programme based on recent work to define components (Kaffenberger et al., 2025).

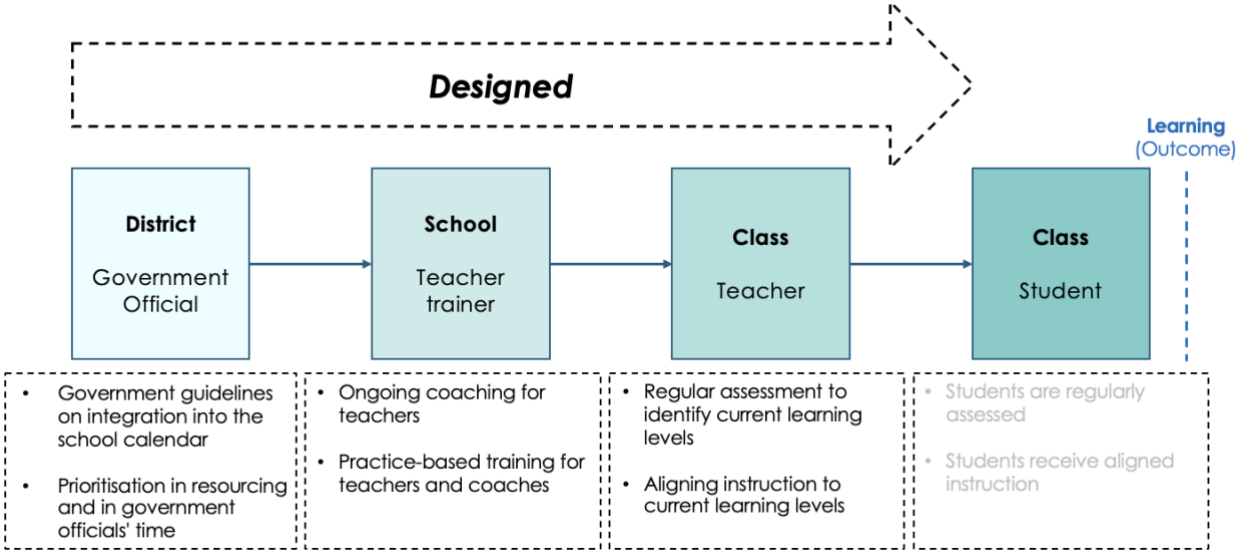


Figure 4: Programme components mapped to implementation stage

Box 2: *Teaching at the right level*

Teaching at the right level is an approach to targeting instruction to children's learning levels in order to help them master foundational literacy and numeracy skills (Banerjee et al., 2017; Angrist & Meager, 2023). It has been shown to be effective in rigorous evaluations across many different contexts in Asia and Sub-Saharan Africa and was identified as a top "Smart Buy" for improving learning by the Global Education Evidence Advisory Panel (Akyeampong et al., 2023). Teaching at the right level programmes typically include frequent assessments; grouping by level; and interactive instruction. A recent paper by Kaffenberger et al., (2025) lays out the "core components" or these programmes.

1.3 What was designed, delivered, and received

Once the existing system and education intervention has been mapped and understood, the next step is to differentiate three key concepts: what was *designed, delivered, and received*. Below we define each of these concepts.

- ***Designed***. Design refers to the implementation plan (i.e. what *should* happen). Components occur at each stage of the implementation chain. This includes defining how the programme components are meant to be delivered at each stage: for example, specifying whether coaching should be provided by district officials and the type of remediation to be offered in the classroom.
- ***Delivered***. Delivery refers to what was delivered in practice (i.e. what *did* happen). Delivery also occurs at multiple stages in the delivery chain. It can be mapped back to the initial design to assess where the programme was delivered according to plan, and where it deviated. For example, remediation might have occurred weekly rather than daily as planned.
- ***Received***. Receipt refers to the final stage of the implementation chain: what did students experience? What was delivered might not necessarily be received. For example, a teacher might conduct a remediation session, but if students do not attend class, they will not receive it.

Interventions often target one or more upstream stages with the expectation that such changes will ultimately enhance student learning downstream. However, as interventions move through the system what is delivered at each stage often diverges from what was originally designed. This can result in the intervention becoming increasingly diluted or distorted, meaning that by the time the intervention reaches students, what they receive is very different from what was planned.

Figure 5 shows an arrow depicting the design of a teacher training programme with the point of intervention starting at the district level, where training takes place. The design arrow runs from the upstream point of intervention all the way downstream to the classroom. The design arrow is outlined with a dotted line to indicate that it is a theoretical possibility not yet implemented in practice.

The solid arrow depicts what was delivered and received in practice, narrowing as it moves downstream. Delivery represents what happens through all stages from the point of intervention. Receipt is the final stage, capturing the student experience (here, in the classroom).

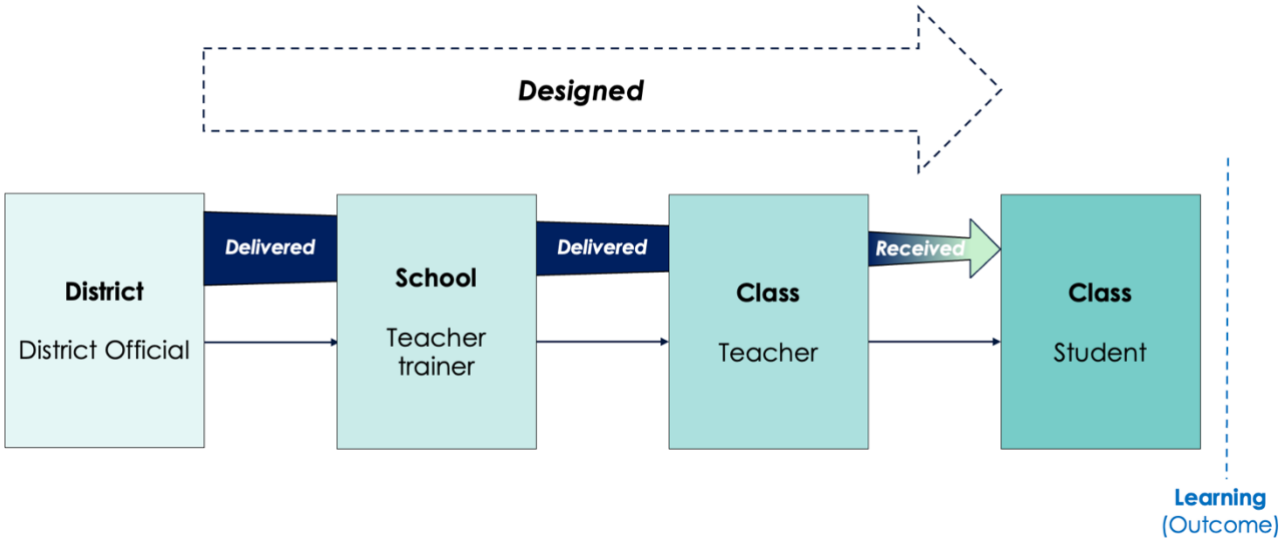


Figure 5: Mapping a simple teacher training intervention onto a system

The final implementation stage, the classroom, has been split into two stages: what is delivered (here, by a teacher) and what is received by the student. While for most stages there is a single agent per stage, the final stage of implementation is segmented into teachers in the classroom and students in the classroom. This additional segmentation in the final stage of implementation is particularly important if there are reasons to believe there are high implementation frictions between the delivery agent and the receiving agent. For example, a lesson may be delivered exactly as intended by the teacher, but students may be absent, disengaged, or unable to understand the language of instruction, meaning that no meaningful intervention was ultimately received.

1.4 Fidelity and take-up

We next formalise the relationship between what was designed and what was delivered or received. We define two concepts: fidelity and take-up.[1]

- ***Fidelity***. Fidelity refers to the relationship between what was designed and what was delivered at a given stage in the implementation chain. Fidelity examines how closely implementation delivery followed the original design. We refer to this type of fidelity as *fidelity to plan*.
- ***Take-up***. Take-up refers to the relationship between what was designed and what was received at the final step in the implementation chain. Only if there is both fidelity and take-up would we expect learning outcomes to improve.

[1] Proctor et al., (2011) propose a related list of 'implementation outcomes': acceptability, adoption, appropriateness, costs, feasibility, fidelity, penetration and sustainability. While each of these outcomes is important, our framework prioritises fidelity and take-up as critical to programme success and more tractable to observe and measure.

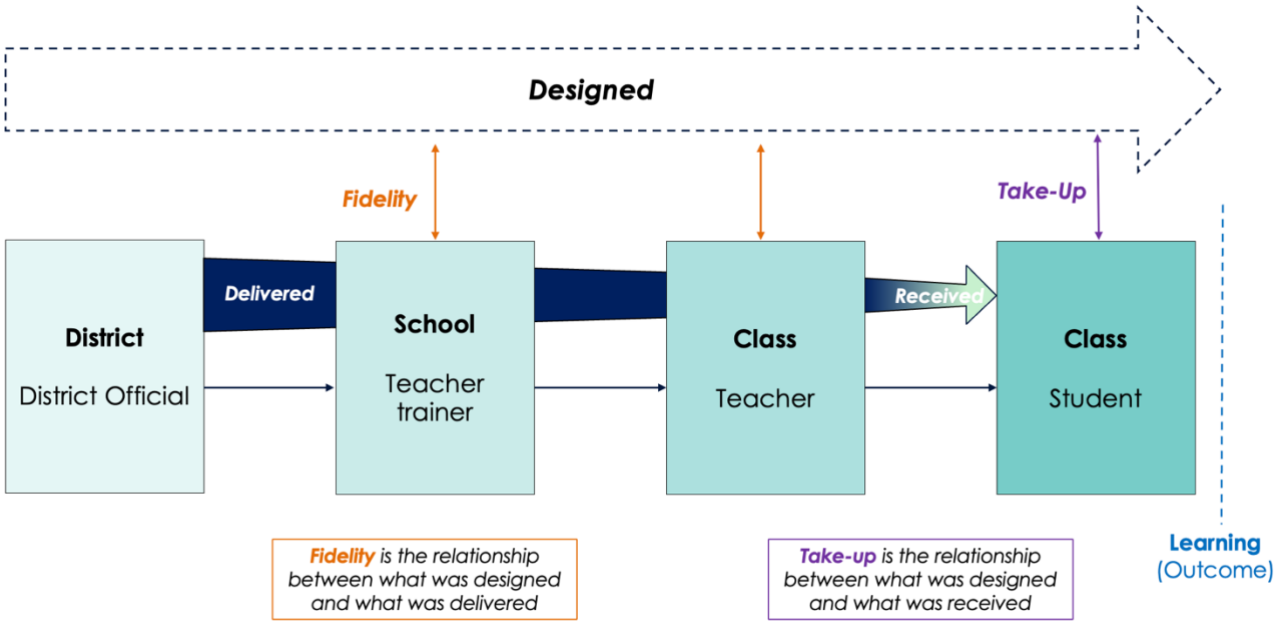


Figure 6: Fidelity and Take-Up

Fidelity metrics can be computed. Since fidelity represents the relationship between what was designed and what was delivered, it can be expressed as a simple ratio:

$$Fidelity = \frac{Delivered}{Designed}$$

Take-up similarly measures the relationship between what was designed and what was ultimately received by students and can also be expressed as a ratio:

$$Take - up = \frac{Received}{Designed}$$

For multi-step interventions, fidelity can be assessed at multiple points along the implementation chain. Generally, fidelity can be measured for as many points as there are stages in the intervention. In contrast, take-up is always computed at the final point of the implementation chain: the student level.[2]

Prioritise one upstream and one downstream stage of implementation to measure fidelity. While implementation can be measured at every stage of an intervention to identify breakdowns as accurately as possible, we recommend a parsimonious approach. Rather than attempting to capture fidelity for every programme component at every stage of implementation, we recommend focusing on priority components at two key fidelity stages in the chain (ideally one upstream and one downstream) where implementation challenges are thought to be most likely.

In addition to measuring two priority fidelity stages, **take-up should be measured wherever possible**, as this is where the intervention's ultimate impact is realised.

[2] 'Student' in this context refers to any potential learner, regardless of whether they are enrolled in school.

Box 3: *The importance of take-up by the final recipient*

The implementation framework defines take-up strictly in terms of what students receive. This is deliberate, and consistent with the view that the ultimate success or failure of any education programme rests with its final impact on students. We note that this may be more challenging to capture for interventions with longer time horizons and/or those aimed at upstream stages in the implementation chain, such as reforms to teacher career pathways. In those cases, we recommend clearly articulating how the final stage of receipt is defined when mapping the intervention and acknowledging what can or cannot be measured.

Box 4: *Understanding the crucial link between fidelity and final outcomes*

While fidelity to plan provides valuable insight into how well implementation is progressing, measuring it in isolation has limitations. Rigid adherence to a plan is not always appropriate; if the plan has flaws or if conditions change, strict adherence can be counterproductive. For this reason, it is important to view measurement as an iterative process, using findings to refine programme design in response to real-world conditions.

Fidelity in this sense is a positive concept, not a normative one. Having high fidelity is not necessarily good or bad in and of itself. Whether high fidelity is desirable depends on the link between implementation indicators and final impact. In each figure above we crucially include the final link to learning, ensuring there is a clear line of sight to how interventions flow through a system all the way to the outcomes of interest. Figure 7 highlights this link and feedback loop.

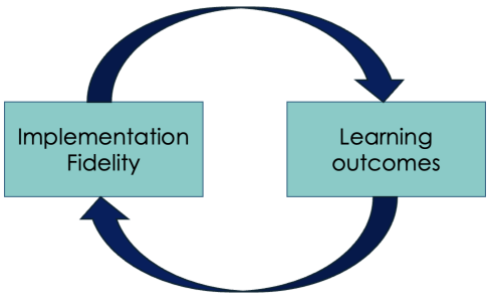


Figure 7: The implementation and learning feedback loop

Table 1 indicates various cases for how implementation fidelity and learning outcomes might correspond with each other and the implications. If implementation is on track, and outcomes are improving, then high fidelity is good. However, if implementation is on track, but outcomes are not improving, it may be better to deviate from and update the plan. Ideally, implementation plans would be updated to pursue better outcomes. A tight link between implementation indicators and learning outcomes is first order to understand if implementation is on track.

Table 1: The decision to deviate from the plan is based on both implementation and outcomes

Implementation proceeding as planned?	Student outcomes improving?	Understanding fidelity to plan: To deviate or not to deviate?

Yes	Yes	**Embrace and optimise the plan:** The programme is on track. Continue measuring implementation and outcomes to maintain impact. Optimise for cost-effectiveness as needed.
Yes	No	**Deviate from the plan:** The programme may be poorly designed or unsuited to the context. Redesign or replace the programme.
No	Yes	**Update the plan:** Agents may be informally adapting the programme to achieve outcomes. Document successful changes to the implementation plan and update the design accordingly.
No	No	**Review the plan:** Diagnose and attempt to address implementation failures. Monitor whether increased fidelity results in improved learning outcomes, or whether a new plan is needed.

Implementation measurement enables implementation research to both *prove* and *improve* – proving whether implementation led to outcomes and improving by enabling real-time course-correction. Ideally, these adjustments occur through **iterative cycles** of measurement and refinement, allowing teams to test changes, check progress, and adapt in real time. For large-scale impact evaluations, this may be addressed through a pilot phase prior to the main programme/evaluation or multiple rounds of iterative or adaptive randomised testing as part of the trial.

For implementers, this can involve regular cycles of iteration and adaptation (see Figure 8). Parsimonious implementation measurement can support iteration and improvement. Selectively measuring a couple key implementation stages enables rapid feedback on whether and at what stage implementation frictions might occur.

Imagine a teacher training programme with a serious implementation flaw: the district officials meant to conduct trainings are overstretched and unclear on the goals of the new programme and thus are slow to roll out the trainings as scheduled. In this case, implementation begins to deviate from design at a very early stage. A parsimonious approach to measurement, focused on one upstream and one downstream stage of implementation, would quickly pick this up. Both measures would show limited fidelity, allowing researchers to quickly diagnose where implementation frictions started to occur and course-correct accordingly.

In this example, the upstream measure might show that very few of the teachers received the full "dosage" of training. This could trigger interviews with district officials to understand why the trainings did not happen as planned. These findings could then inform tweaks to the programme design, for example: streamlining the teacher training materials to focus on the highest-priority messages; reassigning responsibility to for the trainings to a different person; and/or changing the format of the trainings to a hybrid model with some of the material delivered over WhatsApp.

One existing model for this approach is A/B testing (Angrist et al., 2024; Anaman et al., 2026). In A/B testing, randomised, rapid, and regular trials are conducted to quickly compare the effectiveness of different design decisions. To expedite the process and produce insights on a decision-maker's timetable and budget, these A/B tests often run on a rapid schedule (as short as a single school term) and are conducted regularly over time, to ensure continuous improvement. One powerful application of this framework for implementation measurement

would be to pinpoint implementation frictions and identify potential improvements for rapid A/B testing.

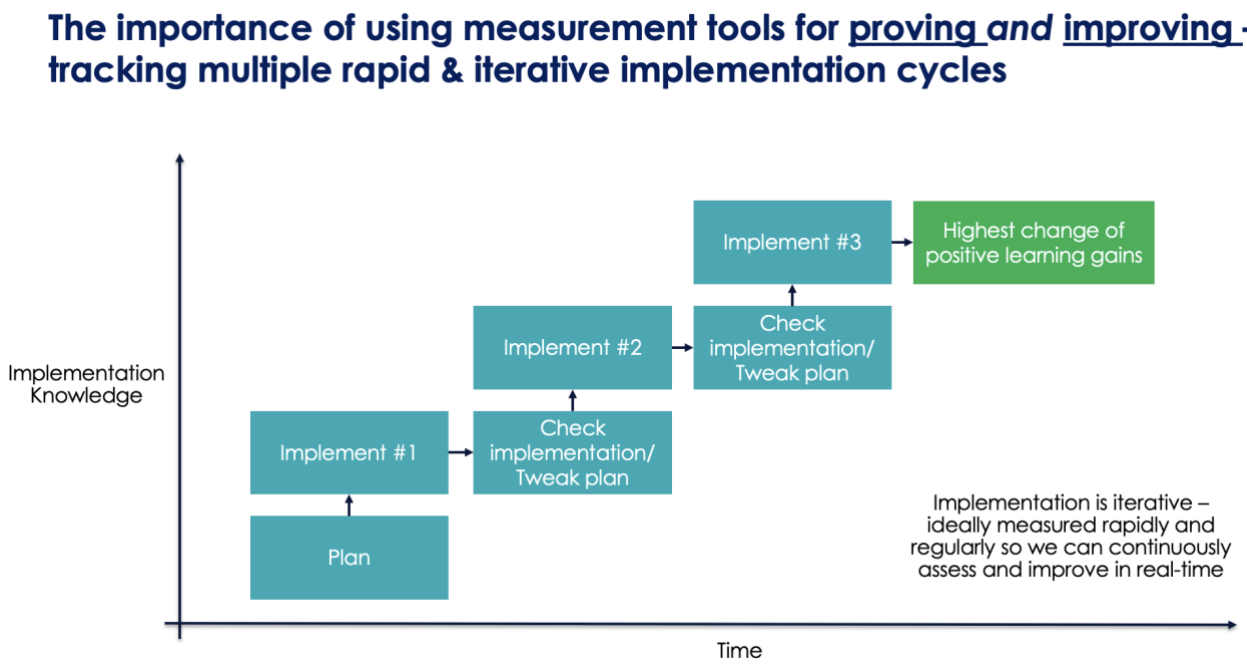


Figure 8: The importance of iteration during implementation

***Box 5:** Defining impact in an education setting*

For the purposes of this note, we define impact in educational programmes specifically as improvements in children's learning outcomes, rather than alternative metrics such as years of schooling or student attendance. This reflects evidence that increased quantity of schooling is not always associated with gains in learning (Pritchett 2013; Angrist et al., 2023). The What Works Hub for Global Education recommends measuring and reporting learning in terms of specific competencies where possible (rather than simply percentages on a test). We are actively developing guidance and resources for learning measurement.

1.5 Fidelity to plan vs fidelity to best practice

Our discussion of fidelity to this point has focused on measuring *fidelity to plan;* an assessment of how closely implementation followed the intervention's original design. For many programmes, especially those that are new or less well understood, this is the most feasible benchmark. In such cases, measuring fidelity to plan is valuable as it helps identify where delivery diverged from expectations and, over time, strengthens the evidence base for future replications of the programme.

For some well-established programmes, however, strong causal evidence exists that links specific programme components to improved learning outcomes. In these cases, implementation can also be measured based on an evidence-based standard – what we refer to as *fidelity to best practice* – rather than only measuring fidelity to plan. This provides an absolute benchmark, not just a relative benchmark, grounded in what is known to generate impact.

Table 2: Fidelity to plan and fidelity to best practice

	Fidelity to plan	Fidelity to best practice
Concept	Relative	Absolute
Relevant to	All programmes	Programmes with well-defined best practices linked to impact

Our conceptualisation of best practice is well-defined. We do not define best practice according to opinion, even when it is expert opinion or sector consensus.

We define best practice as compelling causal evidence linking specific interventions (and their components) to learning. Currently, only a few interventions meet this definition. For example, the 'Smart Buys' identified by the Global Education Evidence Advisory Panel include effective interventions such as teaching at the right level and structured pedagogy that have well-defined components shown to work across contexts (World Bank, 2023). We expect the number of interventions that meet these criteria to grow over time as the evidence base on what works continues to expand (Kaffenberger & Hwa, 2024).

Building on work by Kaffenberger et al., (2026), we use the term "core components" for components that are essential to these programmes' impact. See Box 6 for more details of how core components are defined.

Box 6: *Core components*

Core components are the components essential to a programme's impact. This evidence is derived by triangulating causal evidence from rigorous studies, additional supporting evidence from related studies, and implementer perspectives. For example, Kaffenberger et al., (2026) identify nine core components for teaching at the right level at three levels of the system:

Core components of teaching at the right level		
Policy	**District**	**Class**
Integration into school day calendar	Practice-based training for teachers and coach	Focus on a streamlined set of foundational skills
Prioritisation in resourcing and in government officials' time	Ongoing coaching for teachers	Regular assessment
		Aligning instruction to learning levels
		Interactive instructional techniques
		Low-cost, local and tailored materials

When a plan is fully designed according to best practice, fidelity to the plan will also capture fidelity to best practice. When a plan is not designed according to best practice, fidelity to plan is still valuable to measure. Together, the concepts of fidelity to plan and fidelity to best practice capture not just whether a programme was implemented as intended, but whether it stayed true to the principles that made it effective in the first place.

There can be multiple permutations of fidelity to plan and fidelity to best practice. One can have fidelity to plan, but not to best practice. For example, if a plan indicated that a teacher training should be delivered for a certain number of days, and it was, but the training focused on compliance with new reporting requirements rather than on how to deliver targeted instruction well, this would yield high fidelity to plan but low fidelity to best practice. Conversely, one can have fidelity to best practice but not to a plan. For example, a district official might notice that teachers require additional support and create a local initiative to offer them ongoing coaching. This would be a positive deviation from the original design, resulting in low fidelity to plan but high fidelity to best practice. Outcomes funds explicitly encourage these types of adaptations by paying for outcomes, rather than specific activities or inputs. In these funding schemes, the implementer has the flexibility to innovate over time to improve outcomes. If evidence-based practices indeed improve outcomes, they might be adopted, demonstrating fidelity to a best practice even if they were not part of the original plan.

1.6 Measurement dimensions: Quantity and quality

We have established a set of foundational implementation measurement concepts. Next, we turn to two critical measurement dimensions: **quantity** and **quality**.

- ***Quantity*** refers to the dosage of an intervention. For goods, this typically includes both volume and coverage (e.g., how many textbooks were delivered along with what percentage of intended schools received them). For services, it is a combination of intensity (e.g., hours of training delivered) and coverage (e.g., the proportion of targeted individuals who participated).
- ***Quality*** refers to any changes in understanding or practice at a given stage in the implementation chain.

Distinguishing between quantity and quality indicators is essential because one without the other is unlikely to constitute effective implementation. For example, a teacher training programme that all teachers attend but in which only half understand the content (high quantity with low quality) will be similarly compromised as one where all teachers in attendance fully grasp the material but only 50 out of 100 attend (high quality with low quantity). Total fidelity is the product of quantity and quality, since each element scales the other.

Both quantity and quality can be quantified, a key next step in the implementation framework.

2. Bringing the concepts together in a concrete example

To demonstrate how one might practically apply the framework, we draw on a general teacher training example.

First, we map the intervention to each stage in the implementation chain. Teachers are trained upstream at the district level by district officials and instructed in a new pedagogy. They travel back to schools downstream where they incorporate the new pedagogy in their classrooms, and where students receive it and benefit.

Step 1: *Mapping the intervention to the implementation stages*

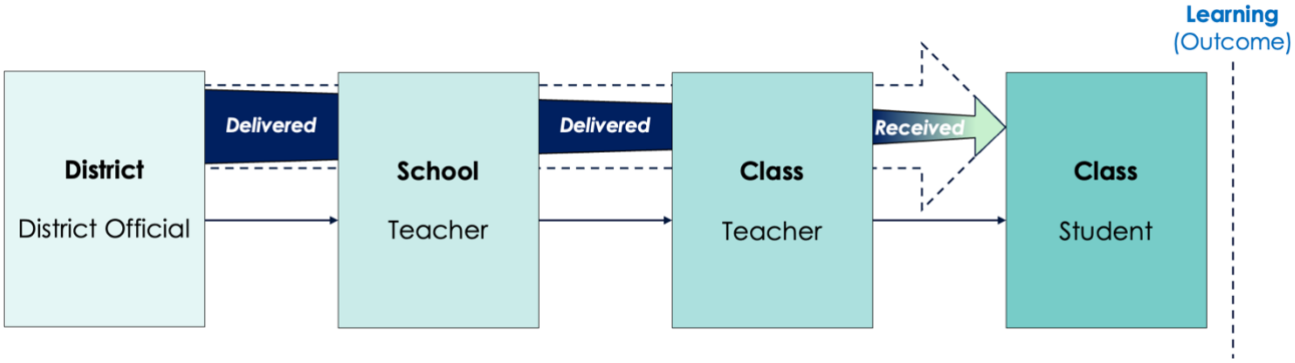


Figure 9: Mapping the intervention to each implementation stage

Second, we specify prioritised programme components, including one upstream component, one downstream component, and a measure of take-up.

Step 2: *Map programme components*

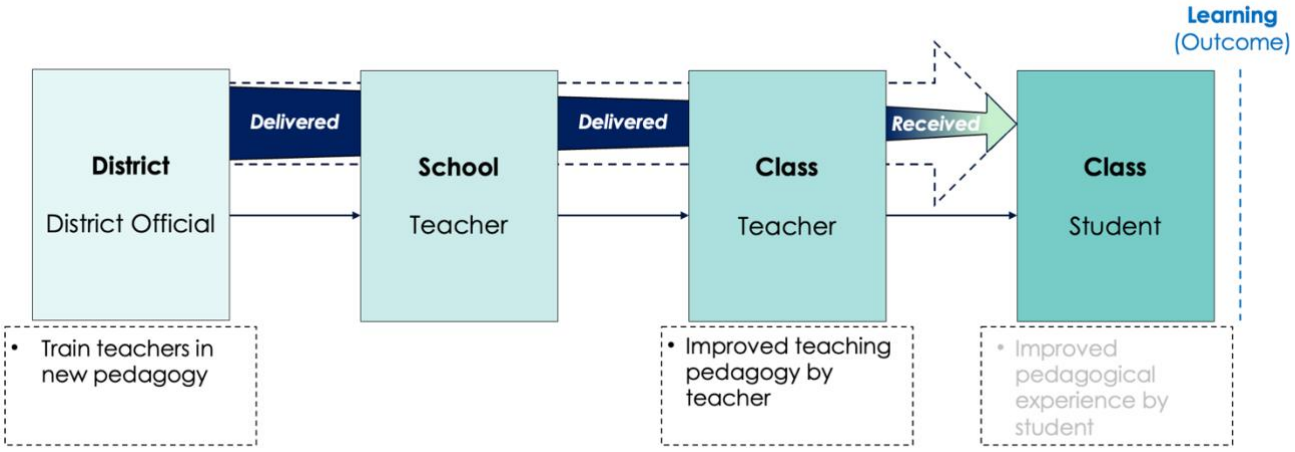


Figure 10: Mapping programme components

Third, we map specific quantity and quality indicators to programme components.

Step 3: *Map programme components to indicators*

Table 3: Map programme components to indicators

Programme Component	Indicator	
	Quantity	Quality
Train teachers in new pedagogy	Number of teachers trained	Percent pedagogical understanding by teachers during training
Improved teaching pedagogy by teacher	Number of teachers teaching new pedagogy in class	Percent of pedagogy delivered in class

Fourth, we define indicators onto implementation stages in the framework: what was designed, delivered, and received. We then compute fidelity and take-up rates. Table 4 includes a 'designed' row, which represents the benchmark for an essential component of the programme, and a 'delivered/received' row, which reflects what occurred in practice. At the bottom of the table, we compute approximate fidelity and take-up rates.

Step 4: *Mapping indicators to implementation stages and computing fidelity and take-up*

Table 4: Mapping quantity and quality indicators

	District Teacher trainer		**Class** Teacher	
	Upstream		Downstream	
	Train teachers in new pedagogy		Improved teaching pedagogy by teacher	
	Quantity	Quality	Quantity	Quality
Designed	100 teachers trained	100% Δ understood pedagogy	100 teachers	100% Δ in classroom practice
Delivered	90 teachers trained	90% Δ understood pedagogy	75 teachers	50% Δ in classroom practice
Fidelity	90.0%	90.0%	75.0%	50.0%
	81.0%		**37.5%**	

	Class Student	
	Downstream	
	Improved pedagogical experience by student	
	Quantity	Quality
Designed	3,000 students	100% Δ in classroom experience
Received	1,500 students	40% Δ in classroom experience
Take-Up	50.0%	40.0%
	20.0%	

Table 4 shows a programme designed to train 100 teachers. These 100 teachers were all meant to adopt a new teaching pedagogy and apply it in their class of 30 students, reaching and benefitting 3,000 students in total.

In terms of upstream delivery, only 90 teachers showed up to the training (quantity) and 90% of them understood the pedagogy (quality). Since the design expected 100 teachers trained with 100% pedagogical transfer, the fidelity rate is 90% in terms of quantity and 90% in terms of quality, with a total fidelity rate of 81% (the product of quantity and quality).

In terms of downstream delivery, we find that 75 of the teachers delivered the new pedagogy in the classroom, and 50% of the pedagogical components were delivered (e.g., frequent diagnostic assessments, but not interactive learning). This results in 75% fidelity in terms of quantity and 50% fidelity in terms of quality, and a total fidelity rate of 37.5%

Finally, in terms of take-up, about 1,500 students received the programme. Even though 75 teachers implemented the pedagogy, a portion of the students in those classrooms were not in attendance. Had all students been present, 2,250 students would have been reached. Moreover, less than half of students experienced improved pedagogy; even though the teacher delivered half of the pedagogical components, a portion of students could not understand the material since they do not speak the language of instruction. The resulting take-up rate is 50% in terms of quantity and 40% in terms of quality for a total take-up rate of 20%.

This example demonstrates how a programme can become attenuated as it flows through the stages of implementation. This illustrates how many implementation challenges over a long implementation chain can compound to produce much lower delivery and receipt relative to what was designed. This also shows how careful implementation conceptualisation and measurement can help identify and address these implementation frictions.

Some of the quality indicators described may appear abstract (e.g., 'change in classroom practice'). This is because the details will depend on the design of the programme being implemented. For example, if key components of the new programme involve both frequent diagnostics and interactive pedagogy, then the relevant measure of quality could be the share of pedagogical components present in the classroom. Differentiating quantity (whether anything was delivered) and quality (the degree of substantive changes) enables a distinction between procedural compliance (or 'tick-box' implementation) versus meaningful changes in practice.

Of note, when making direct comparisons of fidelity rates at various stages of the system, measuring closely related components can facilitate more meaningful comparisons. For example, measuring the degree to which a new pedagogy was effectively communicated to teachers in a training (upstream) along with whether and how teachers implement that pedagogy in the classroom (downstream) could help trace where fidelity to plan is lost and illuminate what parts of the programme eventually reach the student. In contrast, measuring less closely related components, such as coaching upstream versus pedagogical delivery downstream, will not produce a perfectly comparable statistic along the implementation chain; however, it can still reveal more general patterns. Understanding how components relate is critical for interpreting and comparing fidelity and take-up rates.

Programme implementers and researchers should determine appropriate thresholds for fidelity and take-up, considering factors such as cost for the specific context and intervention. For example, a low-cost intervention to share learning content with parents over WhatsApp may be deemed effective if even a small percentage of recipients participate. Conversely, an intensive, in-person tutoring programme during school hours may require close to full participation to justify the cost. Regardless of the thresholds chosen, having plausible benchmarks provides a basis for determining in advance what constitutes suitable levels of implementation, helping to identify which stages might require further attention to improve overall programme performance.

Incorporating measures of fidelity to best practice

We can incorporate notions of fidelity to best practice in our table above (as shown in Table 5). For example, imagine the intervention in question is a teaching at the right level style programme, rather than a generic teacher training. We can now define 'change in classroom practice' more precisely and normatively as whether teachers aligned instruction to learning levels, among other elements – a core component for teaching at the right level to be effective (Banerjee at el. 2017).

Table 5: Dimensions of fidelity to best practice

Programme Component	Indicator	
	Quantity	Quality
Teachers trained on targeted instruction	Number of teachers trained	Percent understanding targeted instruction in training

Teaching instruction aligned to student levels	Number of teachers in class	Change in targeted instruction taught in class

	School Teacher trainer		**Class** Teacher	
	Upstream		Downstream	
	Teachers trained on targeted instruction		Teaching instruction aligned to student levels	
	Quantity	Quality	Quantity	Quality
Designed (according to evidence-based best practice)	100 teachers trained in targeted instruction	100% Δ understood targeted instruction pedagogy	100 teachers	100% Δ in providing targeted instruction by learning level
Delivered	90 teachers trained in targeted instruction	90% Δ understood targeted instruction pedagogy	75 teachers	50% Δ in providing targeted instruction by learning level
Fidelity	90.0%	90.0%	75.0%	50.0%
	81.0%		**37.5%**	

	Class Student	
	Downstream	
	Students received aligned instruction	
	Quantity	Quality
Designed	3,000 students	100% Δ in receiving targeted instruction by learning level
Received	1,500 students	40% Δ in receiving targeted instruction by learning level
Take-Up	50.0%	40.0%
	20.0%	

Box 7 extends the teacher training example by using real data from an evaluation of a teaching at the right level programme in India to show how an intervention can be assessed for fidelity to best practice and improve fidelity to best practice over time.

Box 7: *Fidelity to best practice – lessons from evaluations of teaching at the right level*

Teaching at the right level is a highly cost-effective approach in which students are grouped according to learning level and receive targeted instruction adapted to those levels rather than age or grade. It has been successfully scaled up across South Asia and Sub-Saharan Africa and has helped millions of children to master foundational literacy and numeracy skills. Several influential randomised trial evaluations of teaching at the right level have been conducted (Banerjee et al., 2017). A recent synthesis by the What Works Hub for Global Education identifies a set of core components underlying the programme's effectiveness (Kaffenberger et al., 2026). In this example, we focus on the core component of grouping by learning level and delivering targeted instruction.

However, this core component has not always been delivered with high fidelity. Early results of an evaluation of a teaching at the right level-inspired programme during the school year in Bihar illustrated the consequences of weak implementation. As shown in Table 5, process evaluation data from random school monitoring visits revealed that grouping by learning level and providing targeted instruction were largely absent in the classroom. Consistent with this, the programme produced limited improvements in student learning.

In 2012–13, a strengthened version of teaching at the right level, built on lessons from the earlier trials, was tested during the school year in Haryana. Unlike the earlier model in Bihar, the Haryana design focused explicitly on embedding teaching at the right level into the day-to-day work of teachers and into the routines of the education system. This involved providing practical, hands-on training to teachers, offering ongoing coaching and monitoring, and dedicating an hour in the school day to regroup children by level and deliver targeted instruction. These changes resulted in far more effective implementation. As Table 6 shows, fidelity increased markedly, with grouping by learning level occurring in over 90% of observed classes. These improvements translated into statistically significant gains in student learning, with Hindi language scores increasing by around 0.2 standard deviations.

This sequence of teaching at the right level evaluations illustrates how fidelity to best practice can be operationalised and assessed across contexts and studies.

Table 6: Fidelity to best practice -- process evaluation data in Banerjee et al., (2017)

Indicator	***Delivery model and state in India*** *(% of classrooms that had indicator)*		***Quantity/ Quality***	***Fidelity/ Take-Up***	***Upstream/ Downstream***
	Bihar *(2009-10)*	***Haryana*** *(2012-13)*			
Grouping by level	*0-4%*	*91%*	*Quality*	*Fidelity*	*Downstream*

2. Operationalising the framework through measurement modes, metrics, and tools

2.1 Measurement modes and metrics

Having introduced implementation concepts, we now demonstrate how the framework can offer practical guidance on implementation measurement by informing indicator selection and mode of data collection.

Applying the framework to the example of a teaching at the right level intervention already helped us 1) articulate core components of the programme, 2) map the implementation stages, 3) identify units and agents, 4) distinguish fidelity and take-up, and 5) begin identifying measures of quality and quantity.

This provides a foundation from which to choose specific indicator *metrics* and the appropriate *mode* of data collection. As shown in Figure 11, the ideal approach to measurement is low cost and high credibility. These types of tools can be rolled out at scale and used for implementation measurement consistently and with compelling results. In practice, trade-offs between cost and credibility must often be managed.

- ***Cost***. Implementation measurement can be designed to minimise additional costs to the education system or intervention. For instance, it can be built into existing data collection mechanisms. Costs can also be managed by collecting data at a frequency that enables meaningful ongoing iteration without overloading the system. For mature education interventions, this might mean collecting data once per term or at least once per academic year.
- ***Credibility***. Key dimensions of credibility include construct validity, ensuring consistency among related metrics; predictive validity, ensuring links to eventual outcomes; minimising ceiling or floor effects, and ensuring meaningful variance so indicators capture plausible rates of implementation fidelity.

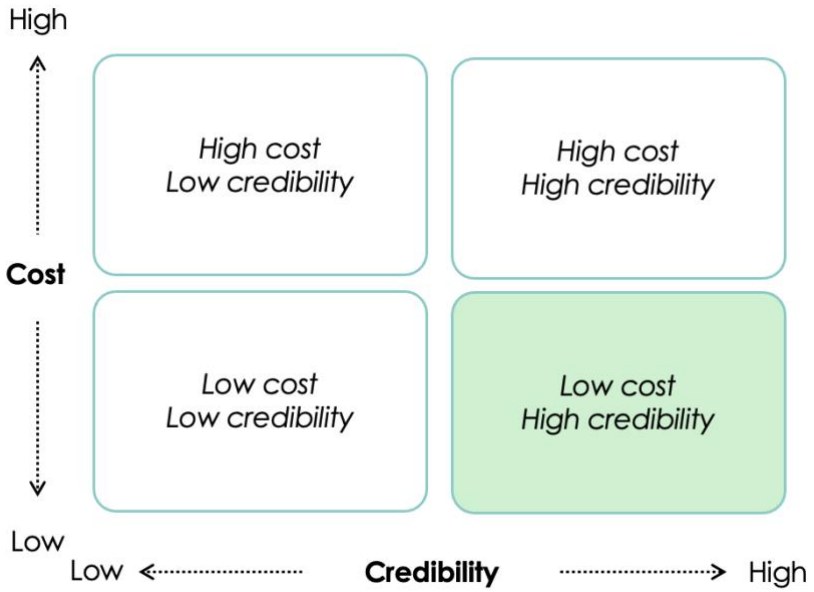


Figure 11: Trade-offs between cost and credibility when choosing metric

Every method has benefits and drawbacks in terms of cost and credibility. In Table 7, we consider four common modes of measurement: in-person surveys, phone surveys, administrative data, and external observations (such as classroom observations).

Table 7: Measurement modes menu

Tool Type	Cost	Credibility	Good for
In-person survey	Medium	High	In-depth, credible information
Phone survey	Low	Medium	Monitoring implementation for real-time improvement
Administrative data	Low	Medium	Monitoring implementation for real-time improvement, when relevant data are collected and reliable data systems exist
External observation (e.g., class observations)	High	High	In-depth, credible information; verifying other modes of data collection

For example, to measure whether instruction is targeted to students' learning levels, one might ask a specific question about whether students appear to be grouped by level in the classroom. Measurement metrics are often bundled with measurement modes, but do not have to be. Whether students are grouped by learning level could be collected through an indicator in a direct survey of teachers, a phone survey of teachers, or through a classroom observation visit from an external monitor.

Triangulating multiple methods can help balance the strengths and weaknesses of each. For example, one might collect information on the prevalence of targeted instruction cheaply at scale through phone surveys of teachers, while conducting classroom observations in a subset of schools to verify the accuracy of those results.

In addition, we encourage following general survey design principles such as those documented in J-PAL's Research Resources (e.g., short recall windows, verification through easily available documentation such as lesson plans and attendance registers, etc.) to increase the accuracy of data collection.

An ongoing research agenda aims to identify the 'sweet spot' of measurement modes and metrics, for example, exploring whether phone surveys can provide high-frequency monitoring over time that can balance rigour with practicality. While classroom observation visits are often considered most credible, they are costly, especially if done repeatedly. If phone surveys of teachers can be similarly credible, they open the door to achieving high credibility at a low cost. The 'sweet spot' lies at the intersection of rigour and practicality (see Figure 12).

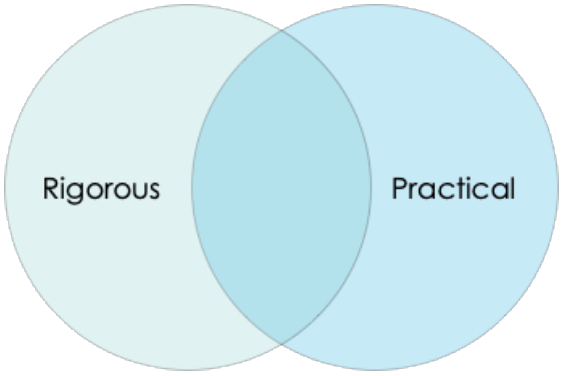


Figure 12: The 'sweet spot' of measurement modes and metrics lies at the intersection of rigour and practicality

We encourage a rich and robust research agenda to explore various measurement modes and metrics, and analysis of which metrics and modes best achieve the sweet spot of low cost and high credibility. The What Works Hub for Global Education is currently working to curate and validate a repository of leading indicators. We also aim to develop and test tools which will be publicly available and continually reviewed and refined.

2.2 Mapping and reviewing existing tools using the framework

When measurement instruments already exist, the implementation framework can be used to strengthen them further. Mapping existing survey or observation items against the framework's dimensions allows researchers to see which aspects of implementation are well covered and where there are gaps. A tool might measure the quantity of delivery (e.g., how many teachers were trained) but overlook quality (e.g., whether teachers understood the concepts from the training). The framework can help identify these omissions and help users to adapt their instruments accordingly before fieldwork. In most cases, this allows existing tools to be strengthened without requiring a complete redesign, which is especially valuable for interventions already being delivered and measured at scale within education systems.

For example, consider the case of teaching at the right level in India and the indicators used to measure implementation in Bihar (Banerjee et al., 2017). This study had unusually rich data on implementation fidelity relative to most of the literature. Random process monitoring visits were conducted at large scale and data was collected on three indicators: (a) teacher attendance at trainings, (b) whether the materials meant to be used in the classroom were indeed present during visits, and, crucially, (c) if students appeared to be grouped by their learning level rather than age or grade, one of the core components of the approach's effectiveness. We conduct a mapping of each of these indicators and how they map onto various concepts in the implementation framework in Table 8. This mapping reveals a few key areas for refinement.

Table 8: Mapping of process monitoring indicators for teaching at the right level relative to classroom core components during the school year in Bihar in Banerjee et al. (2017)

Component	Indicator	Mode	Metric	Quantity/ Quality	Fidelity Take-Up	Upstream/ downstream
Practice-based training	Training attendance	External Observation	90%	Quantity	Fidelity	Upstream
Localised, low-cost, well-aligned instructional materials	Materials present in classroom	External Observation	30%	Quantity	Fidelity	Downstream
Aligning instruction to current learning levels	Grouped by level in classroom	External Observation	4%	Quality	Fidelity	Downstream
Regular assessment to identify current learning levels	-	-	-	-	-	-
Focus on foundational skills	-	-	-	-	-	-

The mapping reveals some strengths and some areas for further consideration. On the positive side, the tool focuses on several core components of teaching at the right level, including both upstream and downstream measures of implementation, and captures measures of both quality and quantity. On the other hand, we see gaps. For example, we see no upstream quality measures (e.g., did teachers understand the material in the training), and the omission of some core components, such as regular diagnostic assessments and a focus on foundational skills, that may be critical to successful implementation. This has practical implications: for example, perhaps the very low rates of targeted instruction are due to poor-quality training, despite high attendance. Moreover, we see no measures of direct student take-up, for example, whether students are *using* the materials which are present in the classroom.

This mapping also reveals opportunities for further refinement in terms of measurement mode. All these indicators are collected via direct classroom observation, a relatively expensive mode of data collection. Testing other modes of measurement could make this type of monitoring cheaper and more sustainable.

We quantify fidelity and take-up using the available indicators against key dimensions of the framework. We use the best indicator available for a given dimension. We show this mapping in Table 9 below. This mapping enables characterisation of fidelity and take-up along the implementation chain. For example, we can see that in terms of fidelity upstream, the intervention in Bihar achieved 90% of its quantity target, but fidelity in terms of quality is unknown. In terms of fidelity downstream, the intervention achieved 4% fidelity in terms of quality relative to the plan and relative to best practice (since grouping facilitates aligning instruction to learning levels, a core component of best practice for the intervention). Fidelity rates in terms of downstream quantity are not directly measured or known. This reveals large gaps between upstream and downstream implementation fidelity, perhaps due to upstream implementation fidelity quality gaps.

Table 9: Fidelity & take-up for one main indicator in Bihar (Banerjee at al. 2017)

	Fidelity		**Fidelity**		**Take-up**	
Unit	***District***		***Classroom***		***Classroom***	
Agent	*Teacher*		*Teacher*		*Student*	
Stage	*Upstream*		*Downstream*		*Downstream*	
Concept	*Quantity*	*Quality*	*Quantity*	*Quality*	*Quantity*	*Quality*
Indicator	*Training attendance*	-	-	*Grouped by level in the classroom*	-	-
Designed	*100%*	-	-	*100%*	-	-
Delivered/ Received	*90%?*	-	-	*4%*	-	-
Fidelity/ Take-up	*90%*	-	-	*4%*	-	-

3. Real-world examples

We now present three use cases to further illustrate how the framework can be practically applied. The first describes an intervention that was implemented unsuccessfully and rigorously evaluated, showing how the framework can help identify where implementation broke down. The second describes an intervention that is currently underway, demonstrating how the framework is being used to track progress and support iterative improvement. The third explores rich measurement, informing high-credibility, low-cost monitoring at scale.

Use case 1: Improving Public Sector management at Scale? Muralidharan & Singh (2020)

In Madhya Pradesh, India, a school quality assurance programme was implemented but failed to produce learning outcomes. The authors go to great lengths to measure implementation and diagnose several implementation bottlenecks. This paper has some of the richest implementation data in the literature and provides an excellent case study. The authors describe the programme as adhering to 'best practice' and being 'implemented well', yet not resulting in improved outcomes.

We apply the implementation measurement framework to the reform and evaluation paper to more precisely pinpoint several implementation frictions.

First, in Figure 13 we map the intervention according to the following implementation stages:

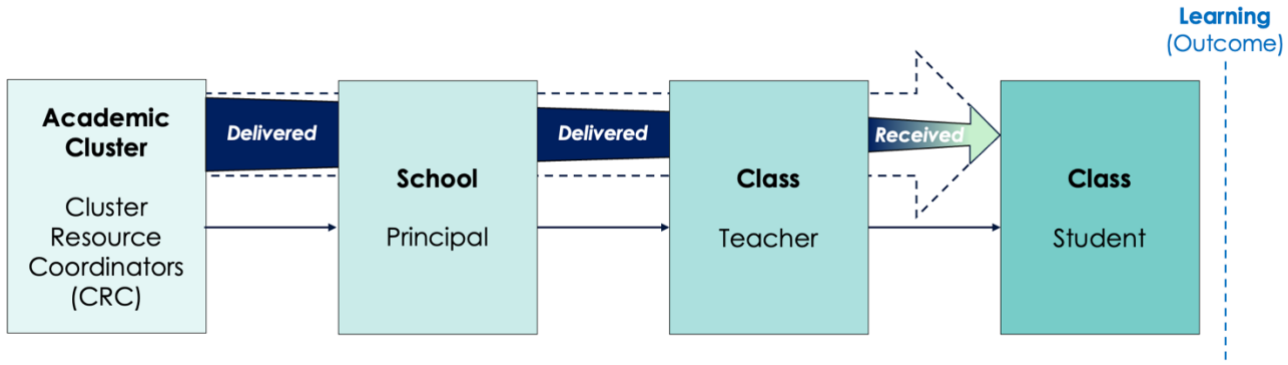


Figure 13: Madhya Pradesh School Quality Assurance Programme implementation phases

Second, in Figure 14 we describe the intervention and several prioritised components at various implementation stages. The intervention focused on school improvement plans (SIPs) and had several components:

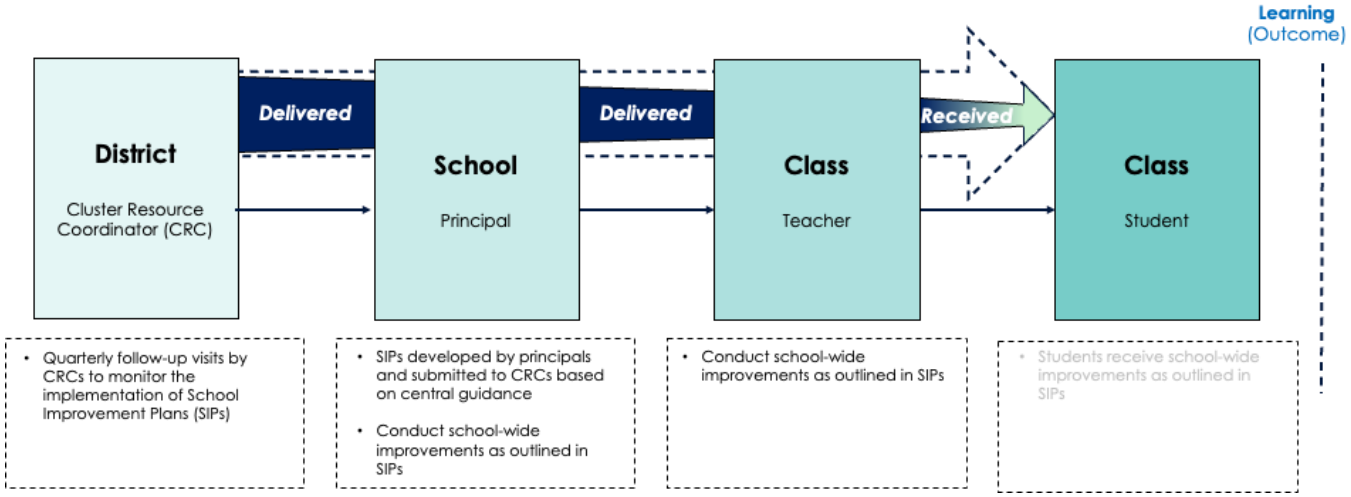


Figure 14: Mapping implementation components for Muralidharan & Singh (2020)

According to the authors, these components were aligned with best practice and delivered successfully. However, the programme failed to increase student learning. Why? Applying our framework helps pinpoint several implementation frictions.

First, the authors describe the programme as adhering to global best practices. However, closer analysis reveals that they define 'best practice' as the popularity of a broad class of interventions appearing in World Bank–funded education projects between 1998 and 2017. This does not align with our use of the term best practice, which defines best practice as specific core components of specific interventions with compelling causal evidence linking them to learning.

This suggests that the intervention may have benefitted from a stronger, more evidence-informed design from the outset. Indeed, a review of the School Improvement Plans (SIPs) reveals that they were long, cumbersome, and unfocused – presenting challenges for busy principals and teachers to implement.

The authors report that SIPs were completed and submitted and that implementation 'happened well.' We assess implementation more precisely using our framework. First, we see that SIPs were filled out and sent from downstream schools to upstream districts. Yet,

while the flow of data from downstream to upstream was reliable, the flow of interventions from upstream to downstream was choppy and poor. We see almost no fidelity in upstream quantity (Cluster Resource Coordinators did not visit schools more) or quality (there are no signs of Cluster Resource Coordinators providing improved coaching feedback). Moreover, we see no fidelity in quantity or quality downstream in the school or classroom (teacher attendance or practice did not change). Finally, we see no take-up at the student-level, either in terms of quantity or quality. Thus, using our framework we would not classify implementation as occurring well, as there were no changes in quantity or quality at any stage of implementation flowing from upstream or downstream. The authors observe the same substantive patterns, referring to this implementation failure as 'administrative compliance.' Using the framework, we characterise the implementation failure similarly, but more precisely: a flow of data upstream, but no fidelity or take-up downstream, neither in quantity nor quality. This leads to *perceived implementation* due to strong incentives to report, rather than effective incentives and systems to deliver real results.

We include a formal mapping in Table 10 below of a subset of implementation stages.

Table 10: Measuring implementation quantity and quality for Muralidharan & Singh (2020)

Unit	**1. Academic Cluster**		**3. Class**	
Agent	**CRC**		**Teacher**	
Stage	Upstream		Downstream	
	Quantity	Quality	Quantity	Quality
Designed	Number of visits	Change in type of coaching feedback	Teacher attendance	Change in type of pedagogical practice
Delivered	No change in number of visits	No change in coaching feedback	No change in teacher attendance	No changes in pedagogical practice

Unit	**4. Class**	
Agent	**Student**	
Stage	Downstream	
	Quantity	Quality
Designed	Student attendance	No quality measured
Received	No change in student attendance	No quality measured

Applying the implementation measurement framework in this way is not intended to imply any shortcomings in the paper. Indeed, the depth of implementation data collected in the study stands out, particularly given that most evaluations devote limited attention to implementation altogether. Rather, the purpose of this exercise is to distil the broader lessons and articulate the rationale for more precisely defining what constitutes 'best practice' and 'well-implemented.' It also underscores the importance of using implementation data not merely for documentation, but for iteratively strengthening programme design and delivery so that interventions pinpoint precise implementation frictions, course-correct, and avoid resulting in administrative compliance to meaningfully reach students to improve learning.

Use case 2: World Bank Implementation Science for Scaling in Education

As a second real world case, we jointly applied the implementation framework to an example where the World Bank Implementation Science for Scaling in Education (ISSE) programme partnered with the Ghana Education Services (GES) and the National Teachers

Council (NTC) to support the Government of Ghana under the World Bank financed Ghana Accountability and Learning Outcomes Project (GALOP, World Bank, 2019). GALOP aims to improve the foundational literacy and numeracy of children in 10,000 of the lowest-performing schools. A central intervention of GALOP is Differentiated Learning, which adapts elements of teaching at the right level to the Ghanaian context. Teachers are expected to group students by learning level, reassess progress every few weeks, and use structured teaching and learning materials to support targeted instruction. They also receive support from School Improvement Support Officers (SISOs), who are expected to provide ongoing coaching and mentoring and make use of standardised tools to record their activities and report data upstream to district and national levels.

As a first step, we mapped the Differentiated Learning intervention and the associated coaching and mentoring model onto the implementation chain, clarifying the sequence of actors, components, delivery mechanisms, and intended outcomes (see Figure 15).

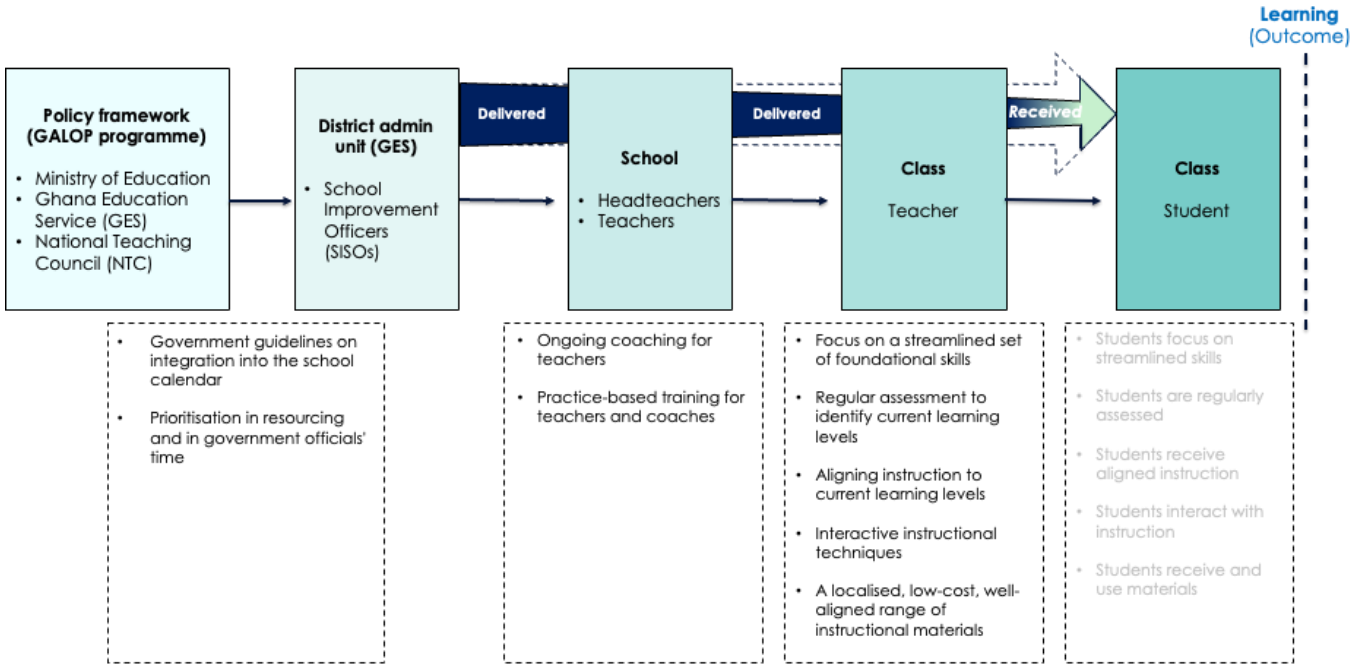


Figure 15: World Bank GALOP project implementation stages

The World Bank ISSE team engaged with GES and NTC to design of a headteacher survey, teacher survey, SISO survey, and a classroom observation tool, drawing on technical inputs informed by the What Works Hub for Global Education implementation measurement framework. Each item across these instruments was mapped against the implementation framework's dimensions and against several core components of teaching at the right level to assess both fidelity to plan and fidelity to best practice. While a broader set of measures was considered during tool design, the final instruments focus on a smaller set of core components, reflecting practical constraints in survey design.

A scalability survey by the World Bank ISSE team identified several areas in which Differentiated Learning was not being implemented as intended. As highlighted in Table 11, three substantive implementation bottlenecks emerged. First, training levels in Differentiated Learning were low. Only a small share of teachers, headteachers, and SISOs reported receiving sufficient preparation to a carry out the programme's core activities (11% to 18%). This suggests a clear upstream quantity bottleneck, which might help explain later gaps in

downstream implementation. Moreover, training quality was not measured, leaving gaps in understanding upstream quality.

Second, although SISO coaches visited schools regularly, the nature of their support differed from the intended coaching model. Most teachers and headteachers reported regular school visits (90% and 100%, respectively). However, the type of support was weighted more towards supervision rather than instructional support. This was reflected not only in what teachers, headteachers and SISOs reported doing, but also in their views about what the SISO role should be. The majority of teachers, headteachers, and SISOs themselves identified supervision as the main aspect of the SISO role (53% of teachers, 82% of headteachers, and 95% of SISOs) while fewer saw coaching and mentoring as central, particularly among teachers (22%, 77% and 94%, respectively). This suggests that SISOs are functioning primarily as supervisors, even though instructional coaching is a core component of the Differentiated Learning programme and of teaching at the right level best practice. This reveals a component with high upstream quantity, but potentially lower quality.

Third, in terms of downstream implementation, student learning levels often went unrecorded and student grouping by learning level was lower than expected. Most teachers and headteachers reported assessing students at least once per term, but fewer than 40% recorded student levels. In practice, this suggests that the Learner Progress Sheets – designed as part of the Differentiated Learning programme to track student learning progress – were not being used as intended. As a result, assessment information was often not carried forward into the next critical implementation step of Differentiated Learning of grouping by level and tailoring instruction for specific learner levels. This highlights a gap in downstream quality.

Altogether, we observe overall impressive rates of implementation along the implementation chain, while also pinpointing several precise implementation gaps, such as (a) increasing the quantity (and likely quality) of training (b) supporting more quality coaching upstream (not just supervision) and (c) improving support downstream, particularly for grouping students by learning level.

Taken together, these implementation insights helped identify implementation bottlenecks and opportunities to test practical modifications to streamline the programme's delivery. Building from these findings, the World Bank ISSE team, in collaboration with government partners and with the technical support of the What Works Hub for Global Education, developed a targeted adaptation to strengthen instructional coaching, including more focus and attention on student grouping by learning level. This took the form of a structured coaching checklist designed to help headteachers provide more consistent and actionable feedback to teachers, with a focus on lesson preparation, peer grouping and tracking learner progress.

The partners collaborated to test the adaptation through a rapid, low-cost A/B test across 150 schools, alongside high-frequency phone surveys to track take-up in real time. Results from the first round show a clear improvement in implementation quality, with the share of classrooms grouping students regularly increasing by around 15 percentage points, at a very low marginal cost (Anaman et al., 2026).

Table 11: GALOP project indicator mapping to the implementation measurement framework

Component	Indicator	Mode and Agent				Quantity/ Quality	Fidelity/ Take-Up	Upstream/ Downstream
		Class observation	*Teacher survey*	*Head teacher survey*	*SISO survey*			

Practice-based training for teachers and coaches	Percentage of teachers who have received sufficient training		*18%*	*14%*	*11%*	Quantity	Fidelity	Upstream
Ongoing coaching for teachers	SISO visited the school or lesson		*90%*	*100%*	*83%*	Quantity	Fidelity	Upstream
	The main role of SISO is to provide monitoring and supervision		*53%*	*82%*	*95%*	Quality	Fidelity	Upstream
	The main role of SISO is to provide coaching and mentoring		*22%*	*77%*	*94%*	Quality	Fidelity	Upstream
Localised, low-cost, well-aligned range of instructional materials	Schools have received materials	*30%*	*88%*	*95%*	*98%*	Quantity	Fidelity	Downstream
Regular assessment to identify learning levels	Assess students at least once per term		*77%*	*87%*		Quality	Fidelity	Downstream
Align instruction by learning level	Group students by learning levels	*44%*	*85%*	*89%*		Quality	Fidelity	Downstream
	Record student assessment score		*39%*	*19%*		Quality	Fidelity	Downstream

Beyond these implementation bottlenecks, the findings also point to broader measurement lessons. Across multiple domains, self-reported implementation systematically overstated what was observed in classrooms. For example, while nearly all teachers and headteachers reported that the school had received teaching and learning materials, classroom observations revealed that only about a third of classrooms had them. A similar contrast appeared for grouping, where more than 80% of teachers and headteachers claimed that students were grouped by learning levels, yet this occurred in fewer than half of the classrooms visited.

This divergence matters because many scalable monitoring systems rely on administrative data or self-reported survey data. Without downstream observation or other validation checks, evaluations risk overstating delivery strength and misdiagnosing where implementation breaks down. These measurement lessons reveal the value of an ongoing learning agenda to identify the "sweet spot" of high credibility, low-cost measurement tools.

Overall, this use case illustrates how intentional implementation measurement, combined with rapid experimentation, can identify and address specific implementation bottlenecks in programme delivery at scale.

Use case 3: Language and Learning Foundation structured pedagogy

The Language and Learning Foundation (LLF) works with state and national government partners in India to strengthen foundational literacy and numeracy at scale. LLF supports implementation through capacity building and system strengthening, including training and field support delivered by government staff and LLF teams. At the classroom level, implementation is tracked through routine process monitoring, including classroom observations of teaching practices and classroom conditions.

Box 8: *Structured pedagogy*

Structured pedagogy is an evidence-based approach to improving instruction quality in low-resource environments by providing teachers with structured lesson plans and ongoing support. Along with teaching at the right level, it was identified as a top "Smart Buy" for improving learning by the Global Education Evidence Advisory Panel (Akyeampong et al., 2023).

In this use case, we applied the implementation measurement framework to review LLF's existing monitoring system and analyse routine process monitoring data to generate implementation measurement insights. The exercise had two aims: first, to identify how current measurements link to the framework's core dimensions (upstream and downstream fidelity and take-up; and quantity versus quality); and second, to identify opportunities to streamline data collection to find the rigour-practicality sweet spot.

As a first step, we mapped LLF's intervention onto the implementation chain, clarifying the relevant units and agents -- for example, middle-tier officials, teachers delivering instruction, and the students ultimately receiving it (see Figure 16).

Second, we mapped LLF's routine monitoring indicators onto the framework's dimensions (see Table 12). This exercise shows that LLF's monitoring system is very rich and sophisticated. At the same time, current measurement is skewed: most indicators capture downstream fidelity, particularly classroom delivery by teachers. Coverage of upstream fidelity, such as support, coaching, and system-level delivery, is more limited, as is direct coverage of student take-up as defined in the framework.

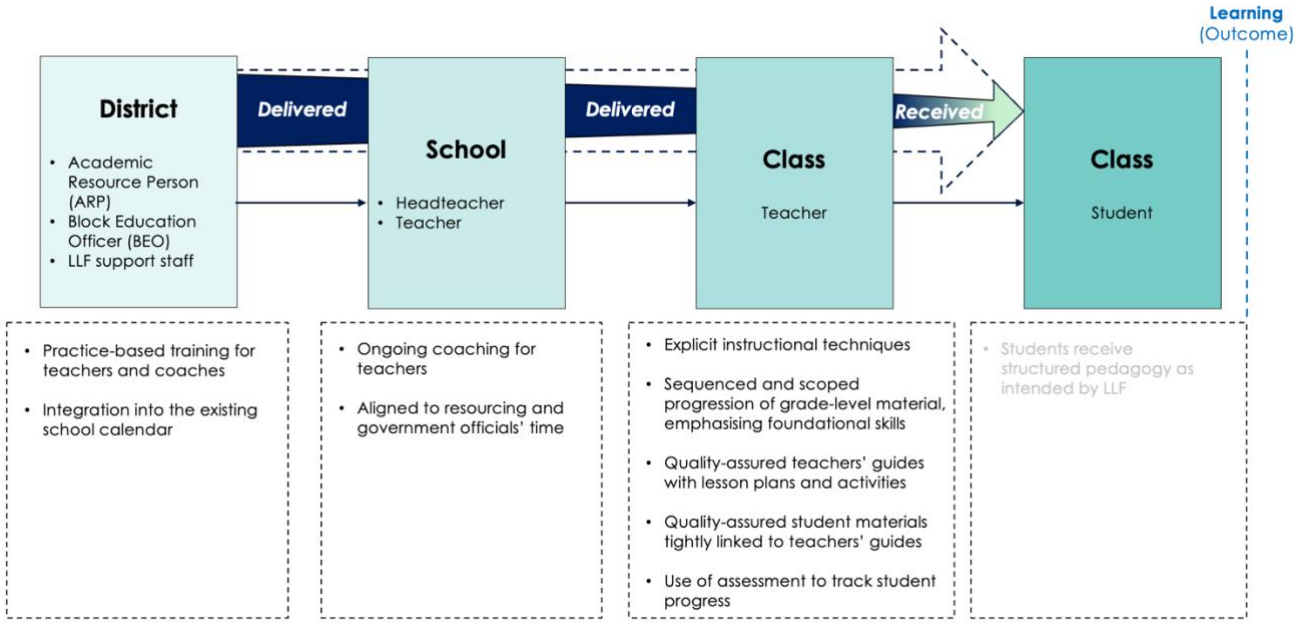


Figure 16: LLF programme implementation stages

Table 12: LLF process monitoring measures mapped to the framework components

Unit	1. School		2. Class		3. Class	
Agent	ARP / LLF team		Teacher		Student	
Stage	Upstream		Downstream		Downstream	
	Quantity	Quality	Quantity	Quality	Quantity	Quality
Designed	No information	100% for all indicators	-	100% for all indicators	100% for all indicators	100% for all indicators
Delivered / Received	• Total duration of school observation	• Demonstration of model lesson • Observers note actionable feedback • Prior feedback reviewed *[Others]*	No data	• Teaching guide used during class • Materials from guide used • Children's understanding checked *[Others]*	• Student attendance • Student participation rate	• Use of learning materials • Access to workbooks • Engaged in activities *[Others]*

Note: For brevity, not all measures are shown. Additional measures are included under "Others".

The mapping also shows that the monitoring tool is heavily weighted toward quality rather than quantity. In the available data, 93% of implementation indicators are quality indicators, while only 3 of the 40 indicators capture quantity. While quality measures may be informative, they are also often more costly to collect reliably at high frequency and may be more vulnerable to measurement error.

We examined how various measurement metrics and modes link to eventual learning outcomes by linking routine classroom monitoring data to student assessments at the school level across 476 matched schools. This analysis indicates which types of indicators are more consistently associated with learning.

A key pattern in the LLF monitoring data is that quantity measures are more strongly associated with learning than the more elaborate set of quality items. The estimated association between reading fluency and the quantity index is about 2.5 to 3 times larger in magnitude than quality (see Figure 17). This suggests that quantity indicators might be a "sweet spot" that is both high credibility and low-cost, facilitating routine measurement at scale. A practical recommendation is to streamline the current set of 40 indicators collected by LLF, most of which currently focus on quality, to a parsimonious subset of informative quantity indicators.

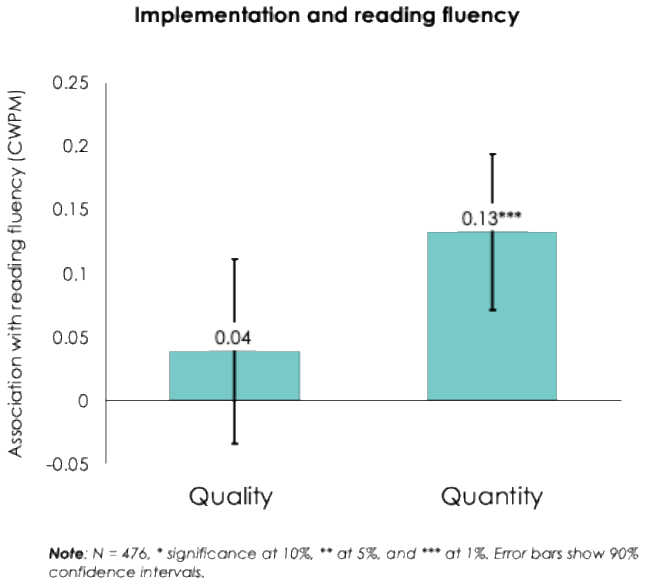


Figure 17: Association between student reading fluency on quality and quantity indicators

This use case highlights how monitoring systems can be strengthened while reporting burden is reduced, and how attention can be focused on a small set of indicators that are feasible to collect at high frequency, at scale, and that are meaningfully associated with learning outcomes.

4. Tools and resources for implementation measurement

The What Works Hub for Global Education is developing several practical resources to support implementation measurement.

Compendium. We are launching an ongoing compendium to curate and share examples of rigorous implementation measurement of education interventions. This is intended as a resource for policymakers, implementers, researchers, and funders, and will inform ongoing refinement and improvement of the implementation measurement framework and measurement tools.

Repository of templates. We will provide templates for mapping education interventions according to the framework to help conceptualise upstream vs downstream stages, component mapping, fidelity vs take-up, and capturing quantity vs quality indicators.

Item bank. We have curated a selection of existing survey instruments from leading researchers and implementers. We are further refining an item bank of indicators relevant to several well-established education interventions. A beta version of the teaching at the right level item bank will soon be available on our website, and we plan to release a version for structured pedagogy interventions in the future. The item bank aims to reduce the burden of creating new instruments from scratch while ensuring that implementation is measured with strong conceptual alignment to the framework.

We further plan to pressure-test indicators to ensure low cost and high credibility, providing researchers with a menu of validated questions that can be adapted to their context. The item bank is being actively piloted and will continue to evolve as new data emerge. Over time, evidence from these pilots will help refine the items, finding a sweet spot between cost

and credibility and identifying which implementation indicators correlate most strongly with learning outcomes. The compendium will also contribute to ongoing item refinement.

These resources will be hosted on the What Works Hub for Global Development website.

Conclusion

The framework we set out can make implementation more visible, comparable, and analytically tractable. Mapping interventions onto the implementation chain clarifies what was designed, what was delivered, and what was received, providing a shared structure for describing how delivery unfolds in practice. The framework helps locate where implementation happens within education systems, clarifying who is responsible at each stage and how these stages connect.

By using consistent categories such as fidelity and take-up, as well as quantity and quality, researchers and implementers can describe different programmes in a common language. Over time, this common structure allows results to be synthesised across studies and contexts, turning individual programme evaluations into part of a broader evidence base not only on *what* works to improve foundational learning, but also on *how* and *why* different approaches can be implemented well to achieve effectiveness.

Our framework does not intend to replace existing theories of change or process evaluations but rather helps organise and connect them. Applied in this way, it offers a guide to interpret variation in outcomes, showing whether differences in learning arise from programme design or from the way delivery unfolded in practice. In doing so, it provides a conceptual basis for linking implementation with impact and for identifying which delivery mechanisms matter most for improving learning outcomes.

Looking ahead, we are embarking on a learning agenda with partners. Much of the field currently relies on bespoke tools. The next phase of work will focus on producing several public goods – streamlined instruments, an accessible item bank, and guidance on reliable metrics that can be used consistently to measure implementation at scale. As these tools develop and are deployed in different settings, we will continue to test and refine them with partners, building evidence on which measures are most predictive of improvements in teaching and learning, as well as how to optimise cost and credibility. The aim is to strengthen the measurement of implementation and implementation itself, identifying 'sweet spot' metrics that balance rigour and practicality.

Finally, we aim to embed implementation measurement within institutional structures. This includes standards in journals, terms of reference provided by donors, working paper sections, pre-analysis and pre-registration plans, and data repositories, among other ideas, to establish norms and support consistent implementation measurement. Strengthening the evidence base on implementation measurement is likely to yield benefits both in the short-run – through more real-time course-correction to keep impact on track – and in the long-run, with implementation becoming a central focus of research, policy, and practice.

References

Akyeampong, K., Andrabi, T., Banerjee, A., Banerji, R., Dynarski, S., Glennerster, R., Grantham-McGregor, S., Muralidharan, K., Piper, B., Ruto, S., Saavedra, J., Schmelkes, S., Yoshikawa, H. (2023). Cost-Effective Approaches to Improve Global Learning - What does recent evidence tell us are "Smart Buys" for improving learning in low- and middle-income countries? London, Washington D.C., New York. FCDO, the World Bank, UNICEF, and USAID. https://geeap.org/wp-content/uploads/2025/10/Cost-Effective-approaches-to-improve-Global-Learning-2023-English.pdf

Anaman, A., Sabarwal, S., Masood, S., Angrist, N. & Spivack, M. (2026). Improving implementation while scaling: Differentiated Learning in Ghana. What Works Hub for Global Education. Insight note. RI_2026/003. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-RI_2026/003

Angrist, N., Beatty, A., Cullen, C., and Matsheng, M. (2024). A/B testing in education: rapid experimentation to optimise programme cost-effectiveness. What Works Hub for Global Education. Insight note, 2024/001. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-RI_2024/001

Angrist, N., Benveniste, L., Bevan, N. & Herbertson, J. (2025). Investing in implementation science, so 'what works' actually works in practice. What Works Hub for Global Education. Blog. 2025/022. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-BL_2025/022

Angrist, N., & Meager, R. (2023). Implementation Matters: Generalizing Treatment Effects in Education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4487496

Angrist, N., Evans, D. K., Filmer, D., Glennerster, R., Rogers, F., & Sabarwal, S. (2023). *How to Improve Education Outcomes Most Efficiently? A Review of the Evidence Using a Unified Metric*. https://doi.org/10.2139/ssrn.4664965

Angrist, N., & Dercon, S. (2024). Mind the gap between education policy and practice. *Nature Human Behaviour*, *8*(12). https://doi.org/10.1038/s41562-024-02013-4

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives*, *31*(4), 73–102. https://doi.org/10.1257/jep.31.4.73

D'Agostino, T., Guzmán, D., Perrin, P., Liberiste-Osirus, A., & Schuenke-Lucien, K. (2024). Explaining Variation in Treatment Effects: An Impact Evaluation and Mixed-Methods Study of Variation in Early Grade Reading Program Effects. *Comparative Education Review, 68(1), 85-112.* https://www.journals.uchicago.edu/doi/abs/10.1086/728393

Hill, C. J., Scher, L., Haimson, J., & Granito, K. (2023). Conducting implementation research in impact studies of education interventions: A guide for researchers. *Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance*, *NCEE 2023-005*.

https://ies.ed.gov/use-work/resource-library/resource/tooltoolkit/conducting-implementation-research-impact-studies-education-interventions-guide-researchers

IPA. (2016). *Goldilocks Toolkit: Monitoring for Learning and Accountability*. https://poverty-action.org/publication/goldilocks-toolkit-monitoring-learning-and-accountability

J-PAL. (2023). Research Resources: Survey Design. https://www.povertyactionlab.org/resource/survey-design

Kaffenberger, M., & Hwa, Y. Y. (2024). A conceptual framework for synthesis and evidence translation to improve implementation of foundational learning. In *What Works Hub for Global Education*. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-WP_2024/003

Kaffenberger, M., Angrist, N., Hwa, Y. Y., Kayton, H. L., Jukes, M., Stern, J. (2026). Core components of teaching at the right level: Unpacking the black box of proven programmes into a set of 'core components' by systematically combining multiple sources of rigorous evidence with implementer insight. In *What Works Hub for Global Education*. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-WP_2026/001

Muralidharan, K., & Singh, A. (2020). Improving Public Sector Management at Scale? Experimental Evidence on School Governance India. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3735697

OECD. (2020). An implementation framework for effective change in schools. *OECD Education Policy Perspectives*, 9. https://doi.org/10.1787/4fd4113f-en

Outhwaite, L. A., Gulliford, A., & Pitchford, N. J. (2020). A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes. *International Journal of Research & Method in Education*, *43*(3), 1–18. https://doi.org/10.1080/1743727x.2019.1657081

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., Hensley, M. (2011). Outcomes for Implementation Research: Conceptual Distinctions, Measurement Challenges, and Research Agenda. *Administration and Policy in Mental Health and Mental Health Services Research,* 38:65–76. https://pmc.ncbi.nlm.nih.gov/articles/PMC3068522/pdf/10488_2010_Article_319.pdf

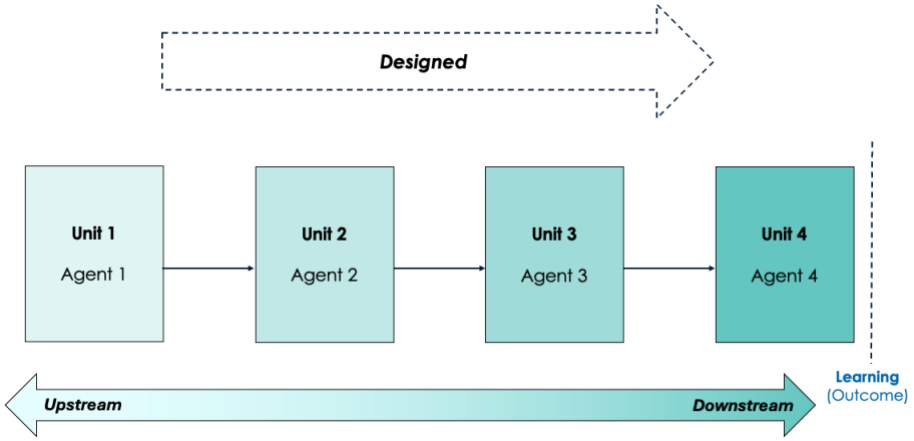
Ryan, A., Prieto-Rodriguez, E., Miller, A., & Gore, J. (2024). What can Implementation Science tell us about scaling interventions in school settings? A scoping review. *Educational Research Review*, *44*, 100620. https://doi.org/10.1016/j.edurev.2024.100620

World Bank. (2019). *Ghana - Ghana Accountability for Learning Outcomes Project (English)*. http://documents.worldbank.org/curated/en/415871570586470453

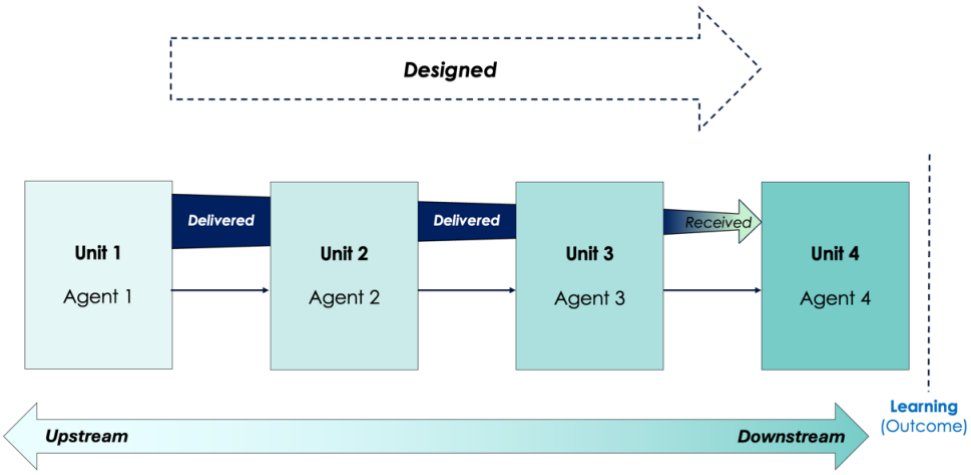
Appendix

Appendix A: Implementation design mapping template

The following figures and tables serve as templates to help research teams delineate the stages, units, and agents targeted by their intervention and map their intervention design/theory of change (see Appendix Figure 1, Appendix Figure 2, and Appendix Table 1).



Appendix Figure 1: Template to map implementation stages in an education system



Appendix Figure 2: Template to map an intervention onto a system

Appendix Table 1: Map programme components to indicators

Programme Component	Indicator	
	Quantity	Quality

Appendix Table 2: Mapping quantity and quality indicators

	Unit Agent		**Unit** Agent	
	Upstream		Downstream	
	Programme Component		Programme Component	
	Quantity	Quality	Quantity	Quality
Designed				
Delivered				
Fidelity				

	Unit Student	
	Downstream	
	Programme Component	
	Quantity	Quality
Designed		
Received		
Take-up		

Appendix B: Implementation tool mapping templates

The following table is a template to help map measurement tools onto the implementation framework. It can be used when adapting an existing tool (e.g., a survey or classroom observation) or designing a new one from scratch

The aim is to assess how well your indicators capture key implementation concepts – fidelity (upstream and downstream), take-up, quantity, and quality.

This table helps classify each component into an indicator, measurement mode, metric and the relevant implementation dimensions. Completing it provides an overview of which aspects of implementation your tool covers and where gaps remain.

Appendix Table 3: Mapping indicators to implementation concepts and dimensions [Template]

Component	*Indicator*	*Mode*	*Metric*	*Quantity/ Quality*	*Fidelity Take-up*	*Upstream/ downstream*

Appendix C: Teaching at the right level core components

Table 4 provides a summary of the identified core components of teaching at the right level. It also highlights the collective strength of the evidence in support of each component. All core components included in the table have been identified to be plausibly essential to the teaching at the right level approach. For more details refer to Kaffenberger at al. (2025).

Appendix Table 4: Teaching at the right level core components

Core component	Essential for effective implementation of teaching at the right level?	Impact evaluations of this component as part of a package	Related programmes and evidence	Implementer perspectives	Summary of the strength the evidence across evidence sources
Core components of the pedagogical programme in the classroom					
Focus on a streamlined set of foundational skills	Essential				Consistently present in effective packages. Substantial additional supporting evidence. Strong implementer backing. Core to the logic.
Regular assessment to identify current learning levels	Essential				Consistently present in effective packages. Substantial additional supporting evidence. Strong implementer backing. Core to the logic.
Aligning instruction to current learning levels	Essential				Consistently present in effective packages. Substantial additional supporting evidence. Strong implementer backing. Core to the logic.
Interactive instructional techniques	Likely essential				Consistently present in effective packages although discussed less often. Some additional supporting evidence. Strong implementer backing.
Localised, low-cost, well-aligned range of instructional materials	Likely essential				Consistently present in effective packages although discussed less often. Some additional supporting evidence. Strong implementer backing.
Core components of pedagogical support					
Ongoing coaching for teachers	Essential				Distinguishing factor between more and less effective packages. Substantial additional supporting evidence. Strong implementer backing.
Practice-based training for teachers and coaches	Essential				Consistently present in effective packages although discussed less often. Substantial additional supporting evidence. Strong implementer backing.
Core components of the authorising environment					
Government guidelines on integration into the school calendar	May be essential				Important in principle but 'how' remains understudied and unclear. Competing evidence from effective but non-integrated holiday camp models. Some additional supporting evidence. Some support from implementer views. Compelling concept, but need for more systematic study of how to operationalise (e.g., centralised or decentralised).
Prioritisation in resourcing and in government officials' time	May be essential				Important in principle but 'how' remains understudied and unclear. Distinguishing factor between more and less effective packages, but tensions around new line items and policies vs tapping into existing ones. Suggestive additional supporting evidence. Some support from implementer views. Compelling concept, but need for more systematic study of how to operationalise (e.g., in-kind resource or new allocation).

Appendix D: Assessing fidelity to best practice

Table 5 provides a simple structure for assessing whether each core component – in this case, teaching at the right level ones – is fully captured, partially captured, minimally captured, or not captured – based on whether the underlying indicators measure the relevant dimensions of implementation in terms of quantity and quality.

- **Fully captured** components are those for which the measurement tool includes indicators that reflect both quantity and quality of implementation in a clear and meaningful way.
- **Partially captured** components are those for which the tool measures both quantity and quality but may not do so comprehensively, or where a small adjustment to an item could allow the component to be captured more fully.
- **Minimally captured** components are those for which only one dimension – either quantity or quality – is represented in the measurement tool, but not both.
- **Not captured** components are those for which the measurement tool does not include any indicators related to the component.

Appendix Table 5: Core components captured – fully, partial, or not captured [Example]

Core Component	Quantity	Quality	Category	Total	%
Practice-based training for teachers and coaches	1	0	Minimally captured	1	3%
Ongoing coaching for teachers	4	2	Fully captured	5	14%
Integration into school day calendar	4	4	Fully captured	8	23%
Focus on a streamlined set of foundational skills	0	0	Not captured	0	0%
Regular assessment	1	0	Minimally captured	2	6%
Aligning instruction to current learning levels	2	5	Fully captured	7	20%
Interactive instructional techniques	3	6	Fully captured	9	26%
Low-cost, local and tailored materials	1	2	Partially captured	3	9%
TOTAL	**14**	**20**		**44**	**100%**

	Number	%
Quantity	14	40%
Quality	20	60%

	Number	%
Fidelity Upstream	14	37%
Fidelity Downstream	20	55%
Take-Up	3	8%

Following from the above table, we would recommend several adjustments to improve coverage of the core components, ensuring the measurement tool captures the elements of targeted instruction most closely linked to fidelity to best practice:

Component	Currently in survey	Suggested addition
Low-cost, local and tailored materials	• Have you received teaching and learning materials for delivering your targeted instruction lessons in your classroom(s)? [quantity]	• How often do you use teaching and learning materials during your targeted instruction lessons? [quantity] • Do you use different teaching and learning materials for different student learning levels during your targeted instruction lessons? [quality]
Practice-based training for teachers and coaches	• Did you attend a targeted instruction intervention training recently? [quantity]	• How many days of targeted instruction training did you attend? [quantity] • How well do you currently understand the core practices involved in delivering targeted instruction lessons? [quality] • Which targeted instruction practices do you feel you understand well enough to apply during your lesson? Select all that apply [quality]
Regular assessment	• How often do you use diagnostic tests to assess students' learning levels? [quantity]	• When you did the diagnostic test, what type of tests did you use to assess student learning levels? [quality]
Focus on a streamlined set of foundational skills	Currently no question captures this component	• How many of your targeted instruction lessons last week focused primarily on foundational skills? [quantity] • Which topics do you typically cover in your DL lessons? [quality]

Glossary of terms

Term	Definition
Agents	Individual or group responsible for delivering an intervention at each stage (e.g., district official, headteacher, teacher).
Core components	The essential elements of an evidence-based intervention, with clear causal evidence showing their link to outcomes. Used to assess fidelity to best practice.
Downstream stages	Later stages in the chain. Namely, the stage that involves the classroom delivery and the student experience.
Fidelity	The degree to which the intervention was implemented as intended. In the framework this is shows as the relationship between what was designed and what was delivered.
Fidelity to best practice	The extent to which an intervention was implemented in alignment with evidence-based core components known to drive impact. Fidelity to best practice provides an absolute benchmark – grounded in causal evidence rather than programme design – showing whether the practices that make an intervention effective were present in delivery.
Fidelity to plan	The extent to which an intervention was implemented as originally designed. Fidelity to plan compares what was delivered (or received) to what the programme designers intended, providing a relative benchmark of implementation.
Implementation stages	Sequential steps through which an intervention moves from policy design to the classroom. Breaking down implementation into stages helps identify where implementation occurs (Units) and who is responsible (Agents).
Indicators	Specific measurable variables used to capture aspects of implementation (e.g., materials present, student grouping).
Measurement metrics	The way an indicator is quantified (e.g., percent, binary, frequency), determining how implementation is measured.
Measurement modes	Methods for collecting implementation data, including in-person surveys, phone surveys, classroom observation, administrative data. Each mode differs in terms of cost and credibility. Ideally tools are low cost and high credibility.
Programme components	The essential elements of an intervention. Used to assess fidelity to plan.
Quality	A measure of how well an activity or component was delivered or received. Quality captures the extent to which delivery reflects the intended practices, understanding, or behaviours – for example, whether teachers understood the training content, whether coaches provided actionable feedback, or whether classroom practices aligned with the pedagogical model.
Quantity	A measure of how much of an activity or component was delivered or received. For goods, quantity refers to volume and coverage (e.g., number of textbooks delivered or percentage of schools receiving them). For

	services, it reflects intensity and coverage (e.g., hours of training delivered and proportion of targeted participants who attended).
Take-up	The extent to which students (or the intended final recipient) receive the intervention. In the framework this is shows as the relationship between what was designed and what was received.
Units	Organisational or physical place at which implementation occurs (e.g., district, school, classroom).
Upstream stages	Earlier stages along the chain. Anything that takes place outside the classroom, we define as upstream.
What was designed	The intended model or delivery plan, including an interventions' activities as well as their expected frequency, dosage, and content.
What was delivered	What implementing agents did in practice, as measured through available data sources.
What was received	The extent to which students (or the intended final recipient) experience the intervention being delivered, as measured through available data sources.

What Works Hub
for Global Education

www.wwhge.org
wwhge@bsg.ox.ac.uk