



# Bringing in Intertemporality




## 2. Methodological Innovations: Bringing in Intertemporality

---

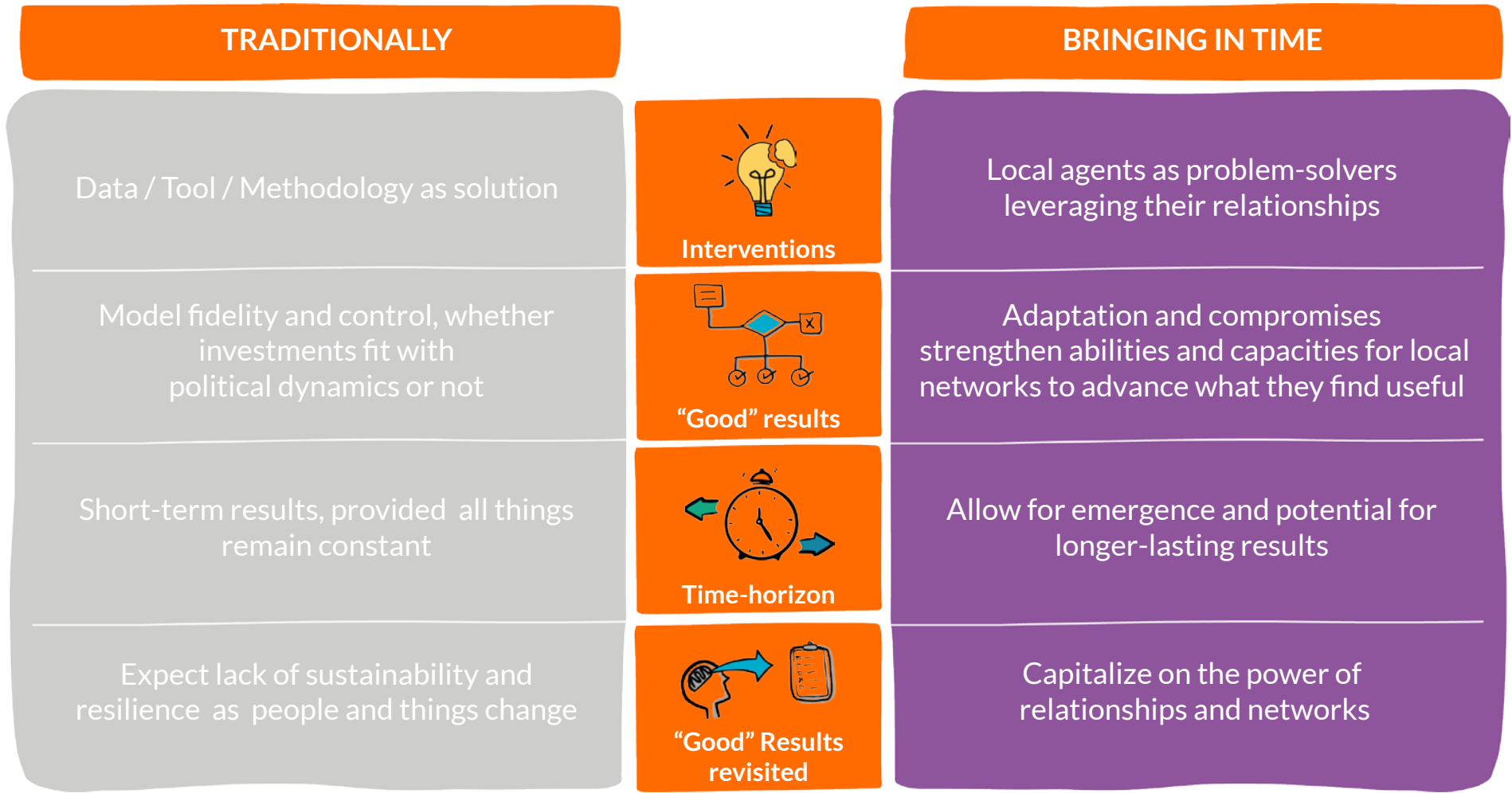
The operationalization of system-aware MEL approaches with causal ambitions can benefit from:

- a) Focusing on relationships as a key driver of systems strengthening;
- b) Zooming in and out of hotspots;
- c) Exploring the (loose) connective tissue between micro-and macro-level change;
- d) Extending time horizons to look beyond individual project cycles (e.g., ex-post evaluation, portfolio evaluation).



This chapter underscores key methodological choices associated with the implementation of a theory-based ex-post evaluation through a “systems-practice” lens. These methodological choices are particularly salient for this evaluation because they enabled us to observe emergent outcomes and understand mechanisms that we might otherwise not have seen (see [Figure 2](#) and read more in [Box 4](#)). For those interested in other, more “conventional” aspects of the methodological framework of the ex-post evaluation, see [Annexes](#).

Figure 2: What are “Smart Buys”?



## 1.1 Transforming Education Systems Sustainably

---

### 1. Spell out your theory-based approach

In **Chapter 1** we set out our approach, discussing:

- a) Why the evaluation took a theory-based approach to systems strengthening, i.e., the assumptions about the nature of the education system it is assessing, how it may become stronger (or not) with reference to the literature on the political economy of education and the assumptions of change-makers, and what may be right tools for evaluators under those conditions;
- b) What specific theoretical assumptions underpin the evaluation and which “causal hotspots” it is most interested in exploring—resonance via layering;
- c) How the evaluators approach and relate to theory-based evaluation for local systems strengthening by combining theory building, refinement, and testing.

### 2. Proactively identify the temporal **boundaries** of the system

What period should be covered by an evaluation? The timeframes of most evaluations are determined, by default, at the starting date of a program, project, or intervention. In particular, the moment when the main implementers of an intervention receive the funds and officially start the work is considered a critical juncture. Indeed, the timeframe typically ends when the project officially winds up and/or the funds run out.

**A system lens calls for looking at the intervention in a (temporal) context.** For most interventions, there are numerous prior decisions that are meaningfully part of the process which constitutes **the intervention** (which includes the people as well as their actions and resources) and, therefore, we should proactively question whether the benefits of assessing them as part of a single story outweighs the added costs.

**The World Vision team made significant decisions before funding was disbursed to the projects** (e.g., selecting partners or key personnel). Furthermore, other (supporting) actors such as donors also made important decisions—from the framing of the intervention to the targeting of regions and schools. Indeed, when these supporting actors made decisions and had either breakthroughs or failures in other projects in their portfolios, or these had direct consequences for the opportunity structure of the projects.

Collectively, as will be discussed in other chapters, these decisions opened or closed doors to the very possibility of layering World Vision’s efforts in specific instances. The resulting blind spot would have been causally significant and biased the assessment.

**Furthermore, we had to be open to listening and triangulating what actors in the system told us were the “right” spatio-temporal boundaries.** Building on [Poli et al., 2020](#), the evaluators had a timeline of possible key milestones dating back more than a decade—long before even the first project, READ. During interviews, however, the evaluators identified a specific event that was not part of our timeline. Had we chosen to leave out this event and of the story by fixing the temporal boundary when the intervention started, or even when the donor decided to open the call for proposals, we would have missed the opportunity to interrogate, observe, and assess whether contributions have been coherent and [added up to more than the sum of their parts](#) over time.

### **3. MEL from different organizational and spatial perspectives**

**Evaluators’ experiences and beliefs shape our understanding of how an intervention or series of interventions may contribute to strengthening a system.** The dynamics we were looking for in the local system might reasonably differ. A single perspective may bias our assessment. For example, building on his previous work, one of the evaluators (Tom) was zooming in. He focused on how layering might have been happening at the micro or school level and how, if at all, approaches, methods, or tools might have trickled up or side-ways across the Dominican education system. Flor, on the other hand, was zooming out. She was focused on layering at a macro and programmatic level, how decisions about the scaffolding trickled down and side-ways, and what discontinuous feedback loops and opportunities for uptake might have been influenced, if not opened, by the loose alignment of actors operating there. As such, she was putting greater emphasis on the long-term trajectory of development partners’ dialogue in the education sector in the Dominican Republic. Associated with these different perspectives there are also slightly different approaches to process-tracing used at different levels of abstraction.

**Secondly, the evaluators chose to look beyond direct contributions within the spatio-temporal boundaries of the two projects studied.** They could have delimited their search to what might be directly attributable to the two projects, as others might. For example, the interview questionnaire used for the [final evaluation of READ](#) suggests that it is unlikely that it could have surfaced insights discussed here (also see [Box 4](#)).

Instead, however, in recognizing that contextual f(actors) can contribute to system strengthening criteria with different shapes—some are straight lines, others may look like a J and still others can look quite different, the evaluators were open to the contributions of actors before the official project boundaries and at adaptations that were made after the projects had ended. When evaluators are looking for a square peg (or linear or J-curve), they can miss the round hole (or the trajectory that has fits and starts, loops, and a different shape).

**Thirdly, the evaluators avoided the lures of organizational-anchoring and credit-claiming, prioritizing the notion that system strengthening is a result that no single organization can achieve on its own.** They considered and valued the contributions of World Vision during, before, and after the two projects had ended. They also explored the role of the two donors that financed and influenced the two projects. Some USAID and World Bank contributions beyond those projects were also considered as part of the overall trajectory of change. More importantly, the contributions of MINERD officials, including national, regional, and district staff, also featured in the analysis. Had they not looked beyond organizational boundaries, the evaluators would have missed the collective nature of the story. This is particularly important in the multiple junctures when staff from one organization move to another that enables (or blocks) the multiplication of insights, cross-learning, opens (or shuts out) opportunities for collaboration, among other functions that are essential for any given organization's effectiveness at a particular moment.

**In complex, nested processes, the two perspectives (zooming in and zooming out) would have been incomplete on their own.** Engaging diverse temporal, organizational, and spatial perspectives enabled the evaluators to understand each level better, questioning each other's prior assumptions, and paving the way for adapting and improving their understanding of the levels of the system. Zooming in and out of nested spheres of influence over a long period also helped the evaluators to grapple more systematically with understanding who, how, when, and why might these levels have been threaded (or not) and what (loose) connective tissue between micro- and macro-level changes might look like.

#### **4. Frame the story in terms of the outcomes that matter most**

Systems-aware social accountability “aims to contribute to a local system that can address problems as they emerge and evolve to respond to new challenges” rather than “for a permanent solution that directly tackles a known problem.” In this context, the evaluators had to reconsider the weight we gave to different system outcomes.

While conventional assessments would have paid most attention to tangible and easily observable outcomes—such as whether an official policy document copied language from a project document and has been adopted, or whether the project’s tool is replicated without being adapted—this evaluation also pays attention to the intangible factors such as relationships, mental models (e.g., new understandings of stakeholders’ roles or what is valued), and power asymmetries that influence whether and under which conditions such language works (or may work) in practice. These intangibles matter in order to understand what it takes for a disengaged parent or guardian to embrace a new role in the school community and participate in action planning, a principal to overcome mistrust of World Vision staff and support the celebration of APMAEs, or for bricklayers to pick and choose which layers they build on and which they discard.

**At its core, this evaluation’s theory of change expects to tell a story that spans multiple project cycles, and one of these intangibles seems to carry much causal weight.** The golden thread is a group of people with relationships doing things based on what they know, who they know, and how they leverage their knowledge and relationships at critical junctures—which may or may not fit with the conventional timelines of individual projects.

**Relationships and the relational infrastructure fuel dynamics in the local system.** They help actors (and evaluators) identify and take advantage of what they believe to be leverage points in the local system. They also matter because they shape change-makers’ experiences, knowledge, and other factors which, in turn, inform how they go about adopting and adapting relevant tools and documents as they navigate changes across the local system. For example, when people implementing a social accountability intervention move to another post within the local system, they may become ambassadors and multipliers for learning from a particular element of an intervention or process implemented in their former job in their new position. When relational dynamics stall, such as when people in government and civil society move to new posts, change-making trajectories often stall (at least temporarily) as a result. When relational dynamics are not included in an evaluation, key aspects of the story are omitted, biasing evaluative judgements.

## 5. (Re)baseline as relevant

When evaluators change the focus from a view of project interventions promoting and decision-makers simply adopting tools at scale to change-agents “co-producing change” embedded in a relational infrastructure, the pre-implementation state in which



they need to test a ToC and assess outcomes also changes. In the case of this evaluation, the team had to reconsider the baseline parameters used for *READ* and *MCPCVME*. Most baselines begin when the contract is signed, thus ignoring the causal significance of previous actions that may have a bearing on project outcomes. For example, while *MCPCVME*'s baseline evaluation (which Tom did) identified numerous potentially relevant causal factors as part of a realist approach, the quantitative data used at the school level to identify the "baseline" came from the endline survey of the *CVME* project to understand things such as whether school actors participated in the election of the school council, for example. It did explain the context-stretching back to the 1997 education law but made only two brief references to the fact that some of the 60 schools had also participated in the USAID-funded *READ* project. The evaluator was indeed unaware that most members of the *MCPCVME* team had worked in the *READ* project (until the mid-term evaluation) and thus did not know about the potential relevance and extent of prior actions and relationships within World Vision. These insights emerged and became salient during this ex-post evaluation. When the evaluation is about change-makers exercising their agency, their prior experiences, and the relational infrastructure in which they are embedded, what really matters is to understand the nature of those factors at the point of establishing the baseline. To observe whether key change agents are part of a group that is able to loosely collaborate, solve problems, and collectively drive change, is only possible when one knows what one is looking for. Only then it is possible to begin observing how the connective tissue of that relational infrastructure (e.g., shared trust, values, experiences, knowledge, goals) is evolving at specific moments and over time.

## 6. Set reasonable expectations

**When evaluators define criteria to assess a process or an intervention, they pre-determine what does or does not count in their evaluative judgements.** Often, evaluations define such limits against an ideal state of an education system—for instance, if 89% of schools in the Dominican Republic have formed APMAES one could aim for 100%, and then expect them to work out fully functional participatory mechanisms that contribute toward every child's quality education. But that kind of change of the education system, as described in **Chapter 1**, has been unrealistic for at least the last 20 years in the Dominican Republic, regardless of the contribution of World Vision's interventions assessed here. Expecting that kind of unrealistic change (i.e., perfect functionality—all parents in every school fully participating, all APMAES formed and working as they are supposed to do) in any education system would pre-define an

intervention as a failure before the evaluation even started to explore the evidence. Instead, “expectations should be tempered given the timescale and resources of a given (project) as well as the nature of progress—often incremental.” Just as importantly, unrealistic expectations can obscure the kinds of learning about plausible, partial, and fuzzy results, which can be more helpful to the intended users of this evaluation and the decisions that they can and do make. It can also lead to hyperbolic tales of triumph and disaster that might do harm to those advancing change.

## 7. Understand how the nature of the process may influence evaluative thinking

**Evaluators’ own thinking about the nature of causal pathways and their interaction with other components in the local system also factor into evaluative judgements.**

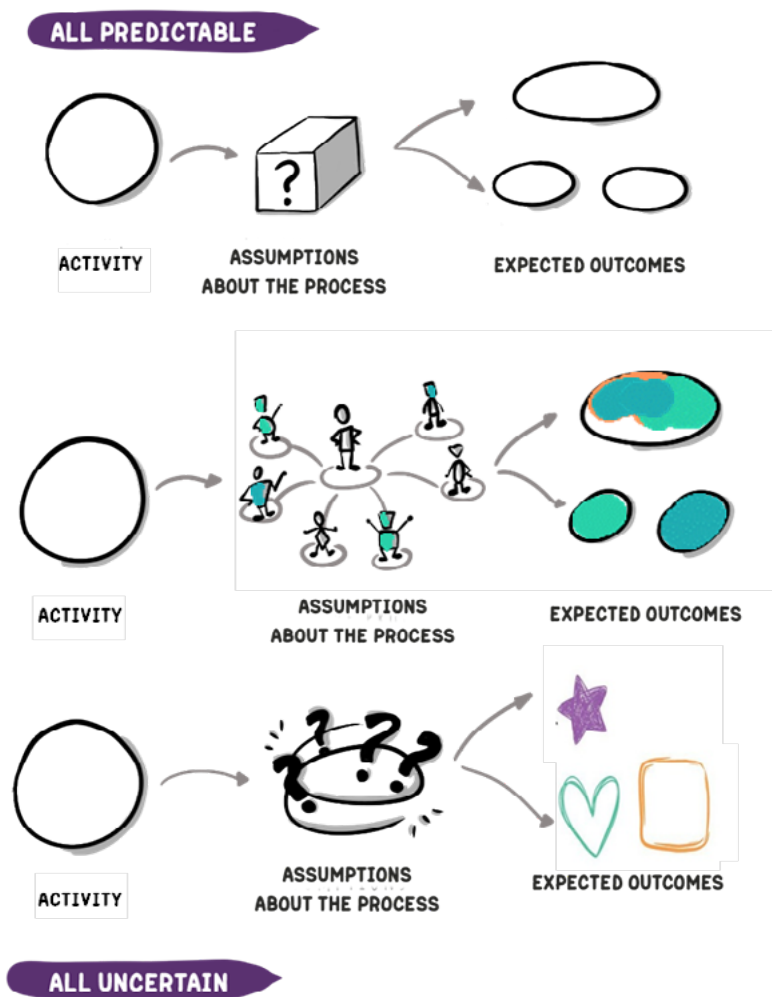
Often, evaluators expect homogeneous results (full reach, wholesale adoption). A good result is one that shows fidelity in the adoption of a tool (e.g., an app), model (e.g., Citizen Voice and Action), language in a document (e.g., use of the term scorecards in Malawi’s National Community Health Strategy, see Box 5) regardless of process and context. At the top of Figure 3, if the intervention looks like a white circle with black borders in, what comes out should also be a white circle with black borders (the process does not matter, which is part of a “black box”).

**In other cases, evaluators may assume that the local system is characterized by uncertainty, and we know a black circle will go in but cannot know or even predict that what will emerge is a purple star, an orange square, or a blue heart—the quest for assessing contributions to systems strengthening is futile (the bottom of Figure 3).** In fact, a good result may be one that is substantively different and fully fits the system’s ever-evolving characteristics.

**In line with these assumptions, the evaluators took a middle route.** We expected good outcomes to have a *partial* continuity with the intervention—whether in function, form, spirit, thrust or other ways—but they also embraced the way the interaction with other actors and elements of the local system might inform emergent adaptations. In fact, adaptations that fit the context might be considered a desirable feature of the process, as is social learning that informs the re-use of bits of the intervention that are most helpful to the actors trying to make new models, approaches, and tools fit with existing ones. As a result of the interaction of process and context, these results are, by design, expected to be heterogeneous, though similarities in the process may produce buckets or types of outcomes that can be meaningfully compared. In the middle of Figure 3 when the multi-colored network

comes into dialogue with the original intervention, some aspects of the original remain in place (in this case the shape), but others change (in this case the colors and sizes). The net of potential outcomes needs to be cast sufficiently wide to avoid overlooking results that should be taken into account, but not so wide that we cannot find identifiable patterns that connect to the interventions.

Figure 3: Comparing expectations of outcomes



Source: Adapted from [Guerzovich and Wadeson \(2024\)](#)


## 8. Evaluate Coherence over Time: Retrospectively and Prospectively

**Local systems strengthening is a moving target.** No matter when it is evaluated, things may continue to grow stronger or become weaker in the future. The movie continues and a snapshot does not do justice to complex processes and outcomes. When evaluators fail to pay attention to the ways in which change-making processes are bound by the past, but also continue in the future, rather than paying attention to the patterns and opportunity structures shaped by time, the quality of their judgments suffers. Furthermore, changes evolve and mutate, so evaluators constantly need to question whether/if this is related to something we did or chose not to do. At some point (with dissipation) it may seem entirely unrelated. But this is easier said than done, especially if the terms of reference for the evaluation do not call for this kind of analysis.

**Retrospective analysis:** There are practical risks to setting out to look backwards, such as potential dissipation of effects. Key informants' and organizational memory loss is also a reality. People move from one job to the next or leave for other reasons—a situation that is particularly acute in education systems where staff, including at school level, change with electoral results—and parents stop engaging in the school when their children graduate. People might only remember parts of the story or remember aspects of it differently than they would have done had they shared it in real time, underplaying, or over-playing events and their and other actors' roles in these. So, looking backwards often requires investing more by critically assessing what informants remember and the documents they can access, as well as triangulating these and various other sources, before reaching an evaluative judgment.

**In addition, one might take steps to focus on cases with the potential for generating data about a longer arc of history or trajectory of change**—a factor that was critical in determining the selection of these two interventions and the specific schools to visit in this evaluation. But many key actors in a project, organization, or school several years ago may no longer still be in the same place at the time of the evaluation, especially if this is many years later. These risks need to be taken alongside risk-mitigation strategies, such as using emergent interviewing strategies and snowballing to find the quantity and quality of relevant sources to make credible analysis and interpretation possible.

**Prospective analysis:** While this is an ex-post evaluation, results are likely to continue to change in the future. According to the OECD-DAC, evaluators can and should think creatively about the future when assessing complex change over time, rather



than making evaluative judgements based on what could be observed at the time when the data were gathered. To do so, it suggests two possible and complementary routes:

- a. Focus on conditions for future outcomes: *Examine if and how opportunities to support the continuation of a process or outcome and its ongoing adaptation have been identified, anticipated, exist and/or have been planned for, as well as any barriers that may hinder the continuation of positive trajectories and/or effects.*
- b. Assess whether processes in the system are on the right track to produce those future outcomes: *Assess how likely it is that any planned or current positive processes and effects will continue, usually assuming that current conditions hold.*

## Conclusion

---

A way to recap this discussion may be to state that evaluating layering in the longer term calls for adding an element of [cathedral thinking to evaluative judgements of processes over a longer period of time \(more than a decade\)](#). Like those laying bricks in a medieval cathedral<sup>13</sup>, bricklayers often have a particular understanding of their place in the temporal context. Those who started building those cathedrals knew they would never see the end of the project in their lifetime, somewhat as many of those who use social accountability to contribute to a stronger education system know that they will never transform systems on their own over the course of a three-year project. The previous generation (or an earlier project) might have laid the groundwork and, ideally, the next ones will take the building (process) further. Often architectural tastes, and insights about what makes education systems perform, also change over the generations (in this case project-based interventions that generally last for three to five years; and presidential periods in the Dominican Republic that last from four to a maximum of eight years). When the newer builders might have preferred a baroque cathedral over the gothic one for which walls had been built, they rarely destroyed what had already been done (and they hope those in the future would not destroy their work). Thus, many cathedrals combine different architectural styles, and in much the same way many education systems reflect trends of different periods.

**The trick to assessing these efforts thus relates to the function (e.g., whether this still a place of worship, or whether this intervention contributes to learning) rather than form.** It is not about a specific brick, tool, document, or personal style, but rather about the effects of the collective rationale and vision that emerge from the various components that people find, along with those they add when exercising their agency. It is not, then, about assessing one person or intervention in isolation but their spatio-temporal interaction within a wider trajectory of sectoral reform. In these cases, evaluation may require a different lens for making judgments, one in which the outcome is assessed in relationship to the abilities of these agents to collectively act, and the [cumulative enabling positive momentum in their desired direction of travel](#).



Bricklayers or stonemasons are similar to makers of an open-ended movie rather than photographers who produce stills. They are anchored in the past, but are building forward. In this open-ended plot change-makers cannot expect all outcomes to be possible as the past takes certain options off the table. Nor can change-makers fully determine how the plot concludes—which is why evaluators and observers cannot fully foresee where the system will “end up” at some future point, leaving questions unanswered. When evaluators are assessing this kind of relational change over time, incorporating “cathedral thinking” into evaluative judgements can support more useful evaluations.



## Box 4

### Smart or best buys?

**In the education sector, the idea of smart buys has become influential.** The basic notion is that it is a straightforward empirical task to point governments and other stakeholders in low-and middle-income countries to interventions elsewhere that are cost-effective in improving learning and education outcomes at scale. As a result, there has been an over-investment in narrow experimental educational research in the Global South, while knowledge about key pivotal political actors and processes is limited.

A recent publication from the Global Education Evidence Advisory Panel convened by the UK Foreign, Commonwealth and Development Office (FCDO), the World Bank, the United Nations Children’s Fund (UNICEF), and the United States Agency for International Development (USAID) classifies “buys” from great to bad, providing a powerful tool to justify decision-making and investments in the education sector.

**The “involving communities in school management” section, which encompasses the kind of interventions covered in this evaluation, is promising, but presents limited evidence.** It notes that providing feedback to schools through community involvement and gathering better data on teachers and students has often had little impact. However, the Panel sustains that where involving community members in school management has worked it has been very cost-effective. It cites five locally bounded short-term interventions assessed through Randomized Control Trials (RCTs) from Gambia, India, Indonesia, and Kenya on what constitutes the evidence.<sup>14</sup> In fact, what the Panel did is simply to cherry pick their preferred studies.

**Firstly, its assessment relies upon a flawed evidence hierarchy which, by design, excludes most of the relevant evidence** in the sector. A wider review of the evidence of 157 interventions, a systematic review of 17 interventions in the sector, and a realist systematic review of 30 interventions in the sector, tell a different story. The effects of social accountability initiatives in the education sector are mixed, but they are broadly positive. But such positive effects are predictably hardest to attain in the most difficult contexts, and are more likely to be achieved when there are various support factors in place.

**Others have argued that the “best buys” approach is a misleading way to assess the true value of interventions.** The best of the “best buys” are those that the Panel most studied in its preferred research methods.



These include teaching teachers how to teach and at the right level. The “best buys” publication itself recognizes that interventions are not the only thing that matters and that systemic reform, which their chosen methods are incapable of assessing, is also crucial. There are serious concerns regarding the lack of external validity of empirical estimates of cost-effectiveness, and indeed the lack of scalability of several preferred interventions (e.g., contract teachers).

**Indeed, it has even been argued that there is a lack of intellectual coherence in making recommendations based on the assumption that all interventions were trying to achieve the same thing** (e.g., improve literacy and numeracy test scores). In these, empirical estimates are themselves part of the evidence that rejects the positive model. Hence, the Panel’s recommendations are speculative, decontextualized, guesswork.

**It is also reasonable to expect mixed results of transaction-intensive (i.e., complex) intervention contexts outside laboratory (or RCT) conditions.** Several replication studies of accountability efforts in the health sector have had different results (positive and negative) because of changes in the underlying contextual conditions in which those interventions were embedded. The very premise of “monocropping” change, which the “best buys” approach exemplifies, works only in conditions where everything else is the same. Once we add change over time, replication studies make little sense. Instead, as has happened in practice in most cases, tools and tactics need to evolve and adapt to respond to the challenges of the day.

**Finally, recent pitches for moving from impact evaluation to implementation research in global education seemed to be based on a similar notion, rather than the one that underpins this evaluation:**

*“a low level of activity on how to respond to real-time implementation challenges using evidence has left international education practitioners with a lack of tested multi-stakeholder models of feedback loops. Without evidence to inform contextualization and adaptation during implementation, donors and governments fund, implement and study the impact of proven solutions again and again but scale with learning has eluded us. This has limited our efficiency in addressing the learning crisis and has delivered more benefits to researchers and donors than children and teachers. We know what education interventions might work at scale but not enough about how to consistently use evidence to adapt and iterate on these solutions to expand their reach with impact. The focus on impact studies has handicapped the use of evidence to explore mechanisms, troubleshoot glitches, and ensure equality in outcomes.”*

The present exercise is an effort to put this into practice.

## Box 5

### An example of uptake

CARE first designed the Community Score Card as a social accountability tool in the Local Initiatives for Health (LIFH) project in Malawi in 2002. It was partly based on community report cards carried out by the Public Affairs Centre in India and participatory rural appraisal (PRA) methods. For several years, CARE Malawi sought to replicate the model faithfully in other projects in the country, and as many as 75 other CARE projects adopted the core of the scoring approach by 2016. CARE USA even copyrighted the term and conducted an RCT of the Maternal Health Alliance Project (MHAP) between 2011 and 2015, which found positive results in various health outcomes such as post-natal and home visits and an increase in the use of modern family planning. However, efforts to scale this up nationally in Malawi's health sector struggled to find financing, but Ntcheu district partially took up a lighter version of the scorecard model at a reduced scale in five rather than 10 facilities between November 2016 and February 2020, and the district development plan mentions scorecards, service charters, and Public Expenditure Tracking Surveys (PETS).