Evaluating large-scale health programmes at a district level in resource-limited countries
Theodore Svoronos & Kedar S Mate

Abstract
Recent experience in evaluating large-scale global health programmes have highlighted the need to consider contextual differences between sites implementing the same intervention. Traditional randomized controlled trials are ill-suited for this purpose, as they are designed to identify whether an intervention works, not how, when and why it works. In this paper we review several evaluation designs that attempt to account for contextual factors that contribute to intervention effectiveness. Using these designs as a base, we propose a set of principles that may help to capture information on context. Finally, we propose a tool, called a driver diagram, traditionally used in implementation that would allow evaluators to systematically monitor changing dynamics in project implementation and identify contextual variation across sites. We describe an implementation-related example from South Africa to underline the strengths of the tool. If used across multiple sites and multiple projects, the resulting driver diagrams could be pooled together to form a generalized theory for how, when and why a widely-used intervention works. Mechanisms similar to the driver diagram are urgently needed to complement existing evaluations of large-scale implementation efforts.

Challenges of evaluation
In January 2010, a retrospective evaluation of the United Nations Children’s Fund’s multi-country Accelerated Child Survival and Development programme was published in the Lancet.\(^1\) The authors found great variation in effectiveness of the programme’s 14 interventions and could not account for the causes of these differences.\(^2\) The journal’s editors wrote that “evaluation must now become the top priority in global health” and called for a revised approach to evaluating...
large-scale programmes to account for contextual variation in timing, intensity and effectiveness.\textsuperscript{3–6}

Evaluations of large-scale public health programmes should not only assess \textit{whether} an intervention works, as randomized designs do, but also \textit{why} and \textit{how} an intervention works. There are three main reasons for this need.

First, challenges in global health lie not in the identification of efficacious interventions, but rather in their effective scale-up.\textsuperscript{7} This requires a nuanced understanding of how implementation varies in different contexts. Context can have greater influence on uptake of an intervention than any pre-specified implementation strategy.\textsuperscript{3} Despite widespread understanding of this, existing evaluation techniques for scale-up of interventions do not prioritize an understanding of context.\textsuperscript{5,7}

Second, health systems are constantly changing, which may influence the uptake of an intervention. To better and more rapidly inform service delivery, ongoing evaluations of effectiveness are needed to provide implementers with real-time continuous feedback on how changing contexts affect outcomes.\textsuperscript{7,8} Summative evaluations that spend years collecting baseline data and report on results years after the conclusion of the intervention are no longer adequate.

Finally, study designs built to evaluate the efficacy of an intervention in a controlled setting are often mistakenly applied to provide definitive rulings on an intervention's effectiveness at a population level.\textsuperscript{9,10} These designs, including the randomized controlled trial (RCT), are primarily capable of assessing an intervention in controlled situations that rarely imitate “real life.” The findings of these studies are often taken, out of their contexts, as proof that an intervention will or will not work on a large-scale. Instead, RCTs should serve as starting points for more comprehensive evaluations that account for contextual variations and link them to population-level health outcomes.\textsuperscript{5,11–18}

The need for new evaluation designs that account for context has long been recognized.\textsuperscript{18–21} Yet designs to evaluate effectiveness at scale are poorly defined, usually lack control groups, and are often disregarded as unsatisfactory or inadequate.\textsuperscript{4} Recent attempts to roll out interventions across wide and varied populations have uncovered two important problems: first, the need for a flexible, contextually sensitive, data-driven approach to implementation and, second, a similarly agile evaluation effort. Numerous authors have proposed novel frameworks
and designs to account for context, though few have been tested on a large scale.\textsuperscript{22–25} Moreover, these frameworks have tended to focus on theories to guide evaluations rather than concrete tools to assist evaluators in identifying and collecting data related to context. In this paper, we review these proposals, present guiding principles for future evaluations and describe a tool that aims to capture contextual differences between health facilities as well as implementation experiences, and may be useful when considering how to best scale up an intervention.

Context-sensitive designs

Several evaluation designs have been proposed in response to the need to understand context in study settings (Table 1). Some of these designs are based on RCTs with changes to allow for greater flexibility. The adaptive RCT design allows for adjustment of study protocols at predetermined times during the study as contextual conditions change.\textsuperscript{14,37} Alternatively, the pragmatic RCT design explicitly seeks to mirror real-world circumstances, especially in selecting participants that accurately reflect the broader demographics of patients impacted by the intervention.\textsuperscript{14,37} Additionally, Hawe et al. propose supplementing RCTs with in-depth qualitative data collection to better understand variations in results.\textsuperscript{22} Each of these approaches has the potential to expand the explanatory reach of the RCT design, and apply its strengths to questions of programme effectiveness and scale-up.

In contrast to alternative RCT designs, theory-based evaluation has been proposed to further understand the actual process of change that an intervention seeks to produce.\textsuperscript{25,38} The most prominent example of theory-based evaluation is Pawson & Tilley’s “realistic evaluation” framework, which is best summarized by the equation “context + mechanism = outcome.”\textsuperscript{17} This framework suggests that the impact (“outcome”) of an intervention is the product of the pathway through which an intervention produces change (its “mechanism”) and how that pathway interacts with the target organization’s existing reality (“context”).

Victora et al. have proposed an “evaluation platform” design that aims to evaluate the impact of large-scale programmes on broad objectives, such as the United Nations Millennium Development Goals. This approach treats the district as the central unit of analysis, and involves the continuous gathering of data from multiple sources which are analysed on a regular basis.\textsuperscript{7} The design begins with the creation of a conceptual model on which data collection and analysis is based, in line with the theory-based approach. The focus on ongoing data collection also
resonates with the work of Alex Rowe, who has advocated for integrated continuous surveys as a means to monitor programme scale-up.\textsuperscript{8}

Alongside these newly proposed evaluation frameworks, some commonly used methodologies have the potential to answer questions of contextual variation. Process evaluations, for example, are increasingly focused on understanding local context rather than simply assessing if each stage of the implementation itself was successful.\textsuperscript{16,22} Process evaluations use several tools and frameworks, including programme impact pathways and results chain evaluations.\textsuperscript{39} Interrupted time series designs also provide an opportunity to understand the effects of sequentially introduced interventions and their interactions with the local environment. These designs have been used in large, multi-pronged studies, in addition to smaller scale applications in conjunction with statistical process control analytic methods.\textsuperscript{40} Multiple case study research also provides a method to study the impact of an intervention on specific individuals (or other units of analysis), allowing researchers to analyse particular drivers behind successful or failed implementation at a local level.\textsuperscript{41,42}

A context-sensitive approach

Each of these approaches attempt to respond to the need to identify and collect local contextual data.\textsuperscript{4,5,7,15,17,20,23} These data will vary significantly depending on the chosen approach, and will be both quantitative and qualitative. Regardless of data type and source, however, the following principles can help guide efforts to capture data on context.

**Standardized and flexible**

Flexibility is an important requirement to successfully capture the role of context, and is also the most difficult to accomplish. It requires developing new qualitative and quantitative approaches, metrics and reliable data collection processes in conjunction with implementers, supervisors and researchers. Choice of metrics will itself be an iterative process that change during data collection. Data collection tools, however, must also maintain a degree of standardization to be comparable across contexts. This is necessary to ensure that implementers understand the “whole” of a large-scale intervention not just its component parts.

**“One level removed”**
A critical question that arises when developing evaluation methods is who will collect and evaluate the data. Potential candidates range from external researchers to the implementers themselves, neither of whom can effectively capture the role of context. An external researcher will have difficulty identifying the situational factors that should be monitored and lacks the intimate knowledge of local context necessary to effectively identify variables across sites. While those actually implementing an intervention will likely possess this knowledge, their perspectives may be subject to multiple biases. We propose identifying an agent who oversees implementation across multiple sites but is still closely involved in implementation activities. This agent would be “one level removed” from the day-to-day activities of programme rollout, thus giving him/her intimate knowledge of the implementation experience without prejudicing the process. Such an agent would facilitate cross-learning and comparisons to produce more generalizable results. Using the district as the unit of analysis, as others have proposed, this individual may be a “supervisor” who visits a subset of clinics as an intervention is rolled out.7

Vetting the data
Despite the particular advantages of a “one-level removed” implementer, the possibility of bias still remains. Key variables on context will need to be validated against multiple sources. Redundancies in currently available data can be used to check on newer data collection tools as they are developed and tested. For example, identifying an inconsistent supply chain as a barrier to implementation could be validated against records of pharmaceutical stocks at facilities. This will be especially true in the early stages of data collection, before the development of formalized structures for collecting contextual data.

A new tool
With these principles in mind, we describe an evaluation tool that aims to capture contextual differences between health facilities and may help programme implementers account for different outcomes for the same intervention in diverse settings. We propose using a specific tool, known as the “driver diagram,” as the central mechanism to capture variation across implementation contexts.43

The driver diagram is a tool commonly used by implementers to understand the key elements that need to be changed to improve delivery of a health intervention in a given context.43 Beginning with the outcome or aim, an implementation team works backward to
identify both the primary levers or “drivers” and the secondary activities needed to lead to that outcome (Fig. 1). Driver diagrams are used in many contexts to assist health system planners to implement change effectively.44–49

In addition to outlining the implementation plan, local facility-based teams develop driver diagrams to help them identify key barriers to implementation and to develop measures to track process improvements to the primary and secondary drivers. The driver diagram can be revisited at predefined times throughout the implementation process, where it is adjusted to account for changes in strategies or unforeseen challenges. The goal of this process is to allow local health system actors to tap into their intimate knowledge of the changing context to more effectively facilitate the implementation process. The iterative nature of the driver diagram process allows adjustments to local context and situation so that, by the end of the implementation, there is a complete picture of that local team's implementation experience.

While driver diagrams have yet to be used specifically for evaluation, they have been widely used to guide implementation in a systematic way. An example is the 20 000+ Partnership, a regional initiative in KwaZulu Natal, South Africa that aims to reduce mother-to-child HIV transmission rates to less than 5%. This project’s initial driver diagram outlined the spectrum of activities that the implementers intended to introduce. On a regular basis, implementers overseeing rollout met to discuss challenges and factors influencing success. These included the introduction of new antiretroviral medications, changes to South African national treatment policies in 2008 and 2010, the launch of a high-profile national HIV testing campaign, changes in local leadership and availability of systems infrastructure (meetings, personnel) to participate in the project.

Each of these meetings provided implementers with an opportunity to understand how local differences between participating sites lead to differences in effectiveness of the intervention activities. Over the course of the project’s implementation, the driver diagrams were modified to reflect ongoing changes.

Though these models were used to guide implementation in this example, there is a clear opportunity for the use of this process in evaluation. The important characteristics that lead to differential outcomes across KwaZulu Natal could provide evaluators with information on
important confounders and effect modifiers, in addition to qualitative data that could contextualize the findings of the evaluation.

Extrapolating from this experience, several teams involved in scaling up an intervention could create local, site-specific driver diagrams and pool these together to show how to best implement that intervention (Fig. 2). This implementation theory would specify consistent findings that could be standardized as well as findings that work best when customized to the local context. Specifically, this product would include: (i) an overall driver diagram reflecting common elements of each facility-specific driver diagram; (ii) a list of diagram components that differed widely across the facility-specific diagrams; and (iii) a list of common contextual factors across the participating organizations’ experiences. This process will allow the public health community to understand the factors that need to be considered when implementing a specific intervention in any given context. Over time, as the driver diagram matures, subsequent implementation efforts would likely be more efficient and more effective.

Taking the concept one step further, if multiple programmes implementing the same interventions around the world pooled their information (and driver diagrams) together, the public health community could develop a more in-depth understanding of that intervention’s dynamics. We join others in proposing an approach similar to the Cochrane Collaboration, in which multiple organizations synthesize their learning in a standardized way. As more context-specific details are fed into this “knowledge bank,” the overall ability to implement interventions at scale will become more accurate, more nuanced and better able to inform future endeavours.

The driver diagram is not without limitations and has not traditionally been used to understand contextual barriers to implementation. Its linear nature is both a shortcoming and an asset as it provides a useful mechanism for organizing complex contextual data but perhaps oversimplifies the same in the process. Its use by programme designers and implementers makes it a useful candidate for bridging the work of implementers and evaluators. While the driver diagram is a useful place to start, we hope that this proposal will catalyse the creation of additional evaluation tools that can capture the role of context as it impacts on population-level health outcomes and draws the implementing and evaluation communities in closer relationship and dialogue.

Conclusion
New models for rapidly implementing efficacious interventions at scale are urgently needed; new ways of understanding their impact are also needed. Through the use of continuous data collection, iterative feedback loops and an acute sensitivity to contextual differences across projects, we can more thoroughly assess the population-level health impacts of interventions already proven to be efficacious in controlled research environments. Further study is needed to develop and test the tools described here in the context of real-time service delivery programmes.

**Acknowledgements:**

The authors thank Gareth Parry, Sheila Leatherman, Don Goldmann, Lloyd Provost and Pierre Barker for their review and comments on earlier versions of this manuscript. Theodore Svoronos is also affiliated with the Mailman School of Public Health at Columbia University (New York, USA), and Kedar Mate is also affiliated with Weill Cornell Medical College (New York, USA).

**Competing interests:**

None declared.

**References**


Table 1. **Overview of context-specific evaluation designs**

<table>
<thead>
<tr>
<th>Evaluation design</th>
<th>Key components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative randomized controlled trials(^{22,26})</td>
<td>Randomized design with flexible protocols to allow for variation, real-world complications and greater external validity.</td>
</tr>
<tr>
<td>Realist evaluation(^{27-29})</td>
<td>Approach designed to understand the interaction between the intervention in question and the context in which it is introduced. Theory building and case study methods are emphasized, though realist evaluation does not rely on a fixed methodology.</td>
</tr>
<tr>
<td>Evaluation platform design(^7)</td>
<td>Aims to assess effectiveness at scale and the contribution of a large-scale programme towards achieving broad health goals. Takes the district as the primary unit of analysis and relies on continuous monitoring of multiple levels of indicators.</td>
</tr>
<tr>
<td>Process evaluation(^{16,30,31})</td>
<td>Assesses the actual implementation of a programme by describing the process of implementation and assessing fidelity to programme design. Relies on various tools to map processes including logic models and programme impact pathways.</td>
</tr>
<tr>
<td>Multiple case study design(^{32-34})</td>
<td>Applies case-study methodology to several subjects with the goal of understanding the complexities of a programme from multiple perspectives. Information is gathered through direct (e.g. interviews and observations) and indirect (e.g. documentation and archival records) means.</td>
</tr>
<tr>
<td>Interrupted time series design(^{35,36})</td>
<td>Uses multiple data points over time, both before and after an intervention, to understand whether an intervention’s effect is significantly different from existing secular trends.</td>
</tr>
</tbody>
</table>
Fig. 1. A basic driver diagram
Fig. 2. A master driver diagram created from an aggregation of project-specific driver diagrams

Generalized theory of change

Better understanding of what works for whom, when and in what context