Centre of Excellence for Development Impact and Learning

MARCH 2022

CEDIL Methods Working Paper 7

Timely evaluation in international development

Marcella Vigneri

LIDC London International Development Centre

UKaid from the British people

Oxford Policy Management

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL methods working papers

The CEDIL methods working paper series offers innovative research methods to develop impact evaluation and evidence synthesis work in low- and middle-income countries.

About this working paper

Please direct any comments or queries to the corresponding author, Marcella Vigneri at marcella.vigneri@lshtm.ac.uk

Suggested citation: Vigneri, M. (2022). 'Timely evaluation in international development'. *CEDIL Methods Working Paper 7*. Oxford: Centre of Excellence for Development Impact and Learning (CEDIL). Available from: https://doi.org/10.51744/CMWP7

Cover design: PhilDoesDesign

Abstract

A central issue in impact evaluation is supporting quick data collection and analyses while an intervention is being rolled out to assist urgent decision-making or update knowledge of what works. This paper reviews approaches to timely evaluation that balance speed with rigour of analysis and are often combined with more standard evaluation methods. We review approaches to timely evaluation from different traditions and combine them in a conceptual framework that describes their goals, speed, and how they address complexity. Each method is paired with a case study to illustrate its value for international development evaluation research.
The authors gratefully acknowledge the inputs of the following colleagues at the London School of Hygiene and Tropical Medicine who were responsible for all the background work and have published a complementary paper on this topic: Jayne Webster, Josephine Exley, James Lewis, James Copestake, Rick Davies, and James Hargreaves. We would also like to acknowledge the invaluable comments of Professor Audrey Prost and Professor Mike Clarke on an earlier version of the paper.

CEDIL methods working paper: Timely evaluation in international development

Author

Marcella Vigneri
London School of Hygiene & Tropical Medicine

Contents

List of tables, figures, and boxes

List of abbreviations

ASPIRES	Accelerating Strategies for Practical Innovations and Research in Economic Strengthening
BCD	Behaviour-Centred Design
BNA	Bottleneck Analysis
CEDIL	Centre of Excellence for Development Impact and Learning – an initiative supported by aid from the UK Government designed to develop
CLARISSA	Child Labour: Action Research-Innovation in South and Southeastern Asia
CSC	Community ScoreCard
DE	Development Evaluation
DFID	Department for International Development (no FCDO)
DHIS2	District Health Information System 2
DRC	Democratic Republic of Congo
EI	Early Intervention
EPI	Expanded Programme of Immunisation
ES	Economic Strengthening
FCDO	Foreign Commonwealth and Development Office
HBV	Hepatitis B Virus
HCC	Hepatocellular Carcinoma
HIV	Human Immunodeficiency Virus
IDS	Institute of Development Studies
IMAM	Integrated Management of Acute Malnutrition
LSHTM	London School of Hygiene and Tropical Medecine
MAM	Moderate Acute Malnutrition
MSC	Most Significant Change
MUVA	An FCDO-funded proramme in Mozambique to help young women in urban areas
mVBR-EI	mobile Village-Based Rehabilitiation – Early Intervention
NSC	National Sanitation Campaign
ODI	Overseas Development Institute
PDIA	Problem-driven iterative adaptation
QCA	Qualitative Comparative Analysis

RCT	Randomised Controlled Trial
RRA	Rapid Rural Appraisal
SAM	Severe Acute Malnutrition
SPC	Statistical Process Control
SWT	Stepped Wedge randomised controlled Trial
UN	The United Nations
UNICEF	The United Nations Children's Fund
USAID	United States Agency for International Development
WHO	World Health Organisation

1. Why do we need timely evaluations?

The motivation for timely evaluations originates from the need to be time sensitive in adjusting the design and implementation of evaluation research as new information is acquired that is relevant for a more accurate analysis of interventions in international development. This requirement is particularly warranted in the evaluation of international development interventions that include multiple activities implemented by different actors. In addition, development interventions are often implemented in contexts with changing features. The interaction between different components of interventions and changing contexts can make their outcomes unpredictable. Theories of change (defined a priori) can rapidly become inadequate when many factors are at play and early assumptions no longer hold.

In such contexts, the use of standard impact evaluation methods may prove of limited value. Randomised controlled trials, for example, require rigid protocols and a long time to test just a limited number of the many hypotheses raised by complex programmes. In complex contexts, there is a need for evaluation methods that can adapt, test different hypotheses, enable ongoing learning, and produce results relatively quickly to inform policy action. Dealing with uncertainty and complexity in international development requires a 'smart and rapid' approach combining flexibility and adaptability with as much scientific rigour as possible within its epistemological paradigm/discipline of origin.

We define timely evaluations as evaluation methods characterised by speed in producing results without sacrificing rigour, and the ability to support adaptation by handling the analysis of different project activities in different and changing contexts. These methods are used before or during an international development programme or intervention in support of its design, adaptation, or refinement. In timely evaluations, data can be collected, analysed, and interpreted to identify necessary, feasible and effective changes at a time when these changes can plausibly improve the programme as needed, and implementers and stakeholders can effectively carry out and benefit from the findings.

In this work we develop a framework that spans different methods for timely evaluations taking place in different contexts and accompany the description of each method with examples of how these methods have been applied to illustrate when they are useful and appropriate. The paper also intends to inform areas of innovation in timely evaluations that CEDIL can promote through its ongoing research agenda. The paper is intended for programme planners and implementers, as well as researchers.

The rest of the paper is organised in four sections. Section two discusses the methodology adopted to review the evidence on approaches to timely evaluation. Section three introduces a framework for categorising approaches to timely evaluation in four categories, depending on the nature of the problem they are trying to address and the goal relative to 'standard' evaluations. Section four reviews the key approaches to timely evaluation in international development, with individual illustration of how they work in practice. Section five concludes with a brief discussion on the approaches reviewed and recommendations on how CEDIL might plan future work in this important area.

2. Research scan/scoping for timely evaluation methods

Our review of timely evaluation methods consisted of four phases: a scoping seminar to set the agenda of the research and identify criteria of timeliness, a public symposium with evaluation practitioners from different development fields using timely evaluations, a literature review scoping timely evaluations methods in the public health sector, and, finally, a broader search for timely evaluation methods and applications in international development.

2.1 Scoping seminar[1]

In June 2017, we organised a scoping seminar at the London School of Hygiene and Tropical Medicine (LSHTM) where members of the Centre for Evaluation (CfE) discussed their own approach to timely evaluations in public health.[2] The event focused on addressing the following five questions:

- What is adaptive learning and how does it differ from quality improvement, performance management?
- How should we define 'timely' data collection (given different context and requirements)?
- How should we define 'timely' data use?
- What processes are used for adaptive learning?
- Is there a role for impact and process evaluations in evaluation of adaptive learning? What differs from other impact and process evaluations?

2.2 Public symposium

Following the reflections from the scoping seminar, the CfE organised a public symposium (November 2017) to include a wider set of approaches to timely evaluation from different disciplines. This public event included 142 participants from the university sector, local and national government, research institutes and not-for-profit organisations.[3]

[1] Section 2.1 to 2.3 draw from earlier work carried out at the London School of hygiene and Tropical Medicine (ins JDEF pub)
[2] See Table A1 in the appendix for a list of participants to this workshop.
[3] See Table A2 in the appendix for a list of participants to the symposium.

2.3 Literature review

We conducted a targeted literature review using a snowballing technique to identify whether adaptive learning approaches in public health had been evaluated (Wohlin, 2014). We began by using the term 'adaptive learning' in PubMed and Web of Science to identify existing evaluations. Additional search terms were added as they were identified. The reference list of relevant literature was screened and developed into an additional forward citation search in Google scholar. The list of search terms is held in Appendix 3.

2.4 Timely evaluations beyond public health

The evidence gathered from the scoping seminar, public symposium and literature review described above focused primarily on timely evaluation methods commonly used in public health. To extend the coverage of our review to other sectors where the issue of timeliness is central, the Centre of Excellence in Impact Evaluation and Learning (CEDIL) extended the review of approaches to timely evaluation to other international development sectors to align the scope of the study with its remit of identifying innovative methods in evaluation research. This additional scoping exercise consisted of three activities: 1. a google scholar search combining the terms "adaptive learning" or "timely evaluation" and potentially relevant sector names such as "education", "employment", "child protection"; 2. a consultation with evaluation practitioners using timely approaches in international development; and 3. a review of methods within the experimental tradition such as factorial designs and adaptive trials and their application to timely evaluations.

3. Framework for approaches to timely evaluation

We first identified four broad approaches to timely evaluation largely based on the disciplines that first launched them. This resulted in a taxonomy of four methodological traditions:

1. Experimental approaches. Although medicine and public health have often focused on the careful design of randomised controlled trials, there is a long tradition of other methods across disciplines for evaluating interventions with multiple activities and outcomes adaptively.
2. Monitoring and quality control approaches. Adaptive methods for testing different activities and outcomes are common in statistical control analysis. These methods have mostly been applied to manufacturing processes with the goal of improving outputs through quality control. Adaptations to these methods are also found in the social sciences.
3. Management approaches. Monitoring methods are used in management, where they are used to test options and make choices in complex contexts.
4. Qualitative approaches. Qualitative approaches can be used to describe and analyse events in complex contexts and offer invaluable granular information that is often essential to inform the direction of adaption.

Table 1 below illustrates the approaches to timely evaluation identified through our review of the evidence, with a classification that reflects the four broad categories defined above:

Table 1: Four categories of timely evaluation approaches

Experimental	Measurement and quality control (monitoring)	Management	Qualitative
Stepped wedge randomised trials	Statistical control	Problem-driven iterative adaptation	Most significant change
A/B testing	Bottleneck analysis	Rapid Cycle Design and Testing	Rapid Rural Appraisal
Adaptive trials	Contribution Analysis	Developmental evaluation	Behaviour Centred Design
Factorial designs	Qualitative Comparative Analysis		Qualitative Impact Assessment Protocol

Below, we provide a brief description of these methods suggesting how and when they may satisfy the following three criteria:

Speed: how quick are these methods?

Goals: what are they meant to achieve?
Fit: what level of project complexity can they address?

3.1 Experimental approaches

A Stepped Wedge randomised controlled Trial (SWTs) design is an evaluation method that includes an initial period in which no sampled cluster[4] is exposed to the intervention. Subsequently, at regular intervals (the 'steps'), one cluster (or a group of clusters) is randomised to cross over from control to intervention status. This process continues until all clusters are exposed to the intervention so that, at the end of the study, all clusters have been exposed to the intervention. Data collection continues throughout the study, so that each cluster contributes observations under both control and intervention observation periods. The SWT approach is used in both explanatory and pragmatic trials. In explanatory designs, decisions on the further implementation of the intervention are left to the end to study to better understand the effect of the intervention. In pragmatic designs, the intervention is primarily offered to generate the intended benefits, rather than to gain research insights. Therefore, decisions about where and when the intervention is to be delivered are mainly driven by practical concerns, although randomisation may be feasible. The method is also implemented when randomisation to either the entire control or treatment arm is not possible; randomisation takes place at a pre-determined date, although it requires adjusting for underlying trends in the data (Hargreaves et al., 2015).

In 1979, The Gambian Government initiated an Expanded Programme of Immunization (EPI) in collaboration with WHO. Under this programme, all children in The Gambia should receive vaccines other than hepatitis B. In West Africa, including The Gambia, nearly everyone is infected with HBV during childhood. For this reason, in 1986, The Gambian government launched as a large-scale vaccination programme for Hepatitis B (HBV) in young infants over a four-year period. Ideally, it is desirable to have evidence of a protective effect of HBV vaccine from controlled trials before launching a mass campaign. However, in the Gambian context, this would have meant delaying the immunisation campaign for 30 years or more. A decision was made to use a stepped wedge trial to ensure that the comparison of HBV vaccinated, and unvaccinated children is not biased by any changes in the risk of children born at different times developing HCC (hepatocellular carcinoma which is known to be strongly associated with HBV infections). This design was considered to be the most appropriate because: 1. immediate universal HBV vaccination was prohibitively expensive; 2. it would be beneficial to have a comparison group of children recruited during the same time

[4] A cluster is a *group* of households, villages, individuals with specific characteristics that are randomly allocated to the intervention arm rather than individual households, villages, people receiving treatment.

period; 3. severe logistic and ethical difficulties would have been encountered with randomisation at the individual level in a trial of this magnitude; and 4. there was hope that the HBV vaccine would be widely available at the end of the study, and that by that time a nationwide delivery system could be in place. The study in The Gambia was considered of critical importance to quantify the precise benefit, if any, derived from mass vaccination with HBV vaccine in early infancy.

Speed: The method is rapid in that it allows to understand important insights of how the intervention delivers its effect as exposure to treatment is sequenced across clusters.
Goal: The goal is to provide research insights in support of (or against) the intervention at the cluster level (though not at the policy level) as the intervention is rolled out.
Fit: Through the sequential transition of clusters from control to intervention conditions in a randomised order, SWTs are particularly useful in the evaluation of complex interventions that take considerable time to influence outcomes.

A/B testing offers a systematic way of finding out what works and what does not in a programme design or marketing campaign. It involves random assignment of participants to different variations of the same intervention. The benefit of the method is that it allows assessments of the effects of adaptations on short-term outcomes. A/B testing is a quick and inexpensive way to test how small changes can have an outsized impact on behaviour. Statistical analysis can then be used to examine which variation performs better for a given goal. The method is particularly useful at the design or pilot stage of a programme to answering earlier on questions on the programme's theory of change.

One illustration of how this method applies in international development interventions comes from the work of Ideas42, a not-for-profit organisation that uses elements of behavioural sciences to address complex social problems. In 2016 the organisation teamed up with JazzCash, one of the largest digital finance and telecommunication multinational companies in Pakistan, to offer digital wallets to allow users to perform a variety of financial transactions (such as sending money to friends and family and paying bills) directly from their phones. In collaboration with JazzCash, Ideas42 designed and tested behaviourally informed text messages to encourage current users to send more referrals to women. The first step was to identify the barriers limiting the take up of new women customers through referrals. Various text messages were subsequently to address these barriers, and test their effectiveness using A/B testing methodology. The test comprised six different messages sent to six different treatment groups of approximately 35,000 users each, which were compared against a group

receiving no messages. Using this A/B testing approach, Ideas42 and JazzCash found that text messages with monetary incentives were most effective at increasing the number of users sending referrals. Prompting clients with social norms and reciprocity messages was also effective, suggesting that social norming, reciprocity, and priming might be sufficient to help more women access these services, even without financial incentives.

Speed: Compares the reaction of users to different methods of intervening in real time.
Goal: Compares two or more versions of a content type to determine which one performs better.
Fit: Method used to run a large number of simultaneous independent tests to compare two versions of an intervention using large volumes of data that can be rapidly collected.

Adaptive trials allow modifications of a trial after initiation without undermining its validity and integrity. It is a method that originates in clinical settings where the goal is to achieve flexibility, efficiency, and speed. Adaptive trials are also often referred to as flexible designs, although the flexibility feature is not loosely defined: the modification and adaptation of one or more aspects of the study design and hypotheses is pre-planned and should be based on the analysis of data (usually interim data) from the study itself. Analysis of accumulating data is then carried out at pre-specified time points during the study. An adaptation is planned before data are examined and consists of a change to either the trial procedure and/or to the statistical procedure during the conduct of the trial. Trial procedures that can be adapted include randomisation method, study endpoints, study hypotheses, sample size, study activities, treatment duration, participant eligibility criteria, criteria for evaluation and assessment of responses, data monitoring and interim analysis. Adaptive design methods often allow the following changes: refining the sample size, dropping treatments cases, changing the allocation ratio of individuals to trial arms, identifying those most likely to benefit and focusing targeting efforts on them, and stopping the whole trial at an early stage for lack of efficacy. All these adaptions are instrumental to ensure that the method become as accurate as possible in capturing information on the intervention that will best produce the knowledge required by the evaluation. Understanding the implications of using an adaptive design, for example, in terms of its practical challenges, and what can (and cannot) be inferred from the results, or how to report and communicate the results are quintessential to the effective use of this approach.

To illustrate how adaptive targeting in a field experiment in development economics, we refer to a job search support programme – Jordan Compact – for refugees in Jordan led by the International Rescue Committee aiming to help Syrian refugees and local jobseekers in Jordan in finding wage work. The field experiment was designed to test three types of support: a small, unconditional cash transfer (worth about one month of average monthly expenditure); information provision to increase the visibility of skills to employers; and a behavioural nudge to strengthen job search motivation. The impacts of these interventions were measured through three follow-up surveys, all administered by phone. Participants' employment outcomes were tracked with a 'rapid follow-up' phone interview six weeks after treatment simply asking about wage employment. The researchers then run detailed follow-up surveys two and four months after treatment to measure a broader set of impacts and to study the effects of the intervention over a longer time. The use of an adaptive design testing different permutations of interventions iteratively achieved better treatment outcomes than a standard parallel RCT design while also learning about each treatment arm with high precision (Caria et al. 2020).

Speed: Rapid adaptation within the timeframe of a trial.
Goal: Learn to adapt the design quickly based on interim data, evidence, and short-term impact.
Fit: When researchers/programme planners need to adjust an intervention while it is ongoing to maximise its potential effectiveness.

Factorial designs are useful when researchers and/or programme planners are interested in looking at different interventions either on their own or as combinations, to determine which works best. A 'factor' is designed as a major independent variable, some typology of treatment, whereas level is a subdivision of a factor. The number of different treatment groups that are considered in any factorial design is determined by multiplying the number of possible factors by the total number of factor levels across all variables. Factorial designs have several important advantages, the main one being flexibility in exploring or enhancing the "signal" (treatment) in any given study. Whenever there is an interest in examining treatment variations, factorial designs are considered an attractive design option because they allow investigators to test different intervention permutations in one rather than conducting a series of independent studies. The method is also considered to be effective in examining interaction effects.

A 2018 study from the Institute of Development Studies (IDS) provides an illustration of what the method offers. The FHI 360 project 'Accelerating Strategies for Practical

Innovation and Research in Economic Strengthening' (ASPIRES) provided technical assistance to a youth programme in Pretoria, South Africa, to assesses if integrating an HIV-prevention education intervention with an economic strengthening intervention improved both economic and health outcomes beyond individual interventions. The youth selected were randomised into four groups: combined economic strengthening and HIV-prevention interventions; economic strengthening intervention only; HIV-prevention education intervention only; or no interventions. The study used a full 2x2 factorial design and randomly assigned participants to ES, HIV, ES + HIV, or control. A linear model would have estimated the effect of each intervention and the effect of the integrating them by using interaction terms. However, the linear model cannot test whether integrating different versions of the intervention is more effective in addition to the single intervention effect. A full factorial design enables measurement of both the impact from integration and from synergy: 1 + 1 > 2 (synergy and amplifying effects) and 1 + 1 > 1 (integration effects where in the absence of synergies the combined effect remains higher that individual effects alone).

Speed: Can be used as a rapid screen to identify factors that are of most importance in determining response to an intervention.
Goal: To discover which factors influence the outcome of the experiment and what levels of these factors lead to an experiment with the greatest sensitivity.
Fit: To look at different interventions either on their own or as combinations, to determine which works best.

3.2 Monitoring methods

Statistical Process Control (SPC) is a method originally created as an industrial tool for measuring the performance of routine manufacturing processes. It subsequently developed into a statistical evaluation tool to help data users make logical decisions regarding the need for change in procedures related to an evaluation. The underlying assumption of SPC is that if the occurrence of a particular event is examined over time, any departure from the expected distribution of the event will be used as information to update and improve the process. SPC uses chronological data charts to overcome the impracticalities of frequent quality control data, which typically show statistically significant but not necessarily relevant changes in processes. By looking at performance over time, SPC makes variation in the process explicit, shows what the process is current capable of delivering, and helps discern between special and common causes of variation.

To illustrate how the method is used in international development, we draw from a healthcare intervention in Syria, which aimed at reducing Ventilator-Associated

Pneumonia (VAP), one of the most common infections in patients requiring endotracheal tubes with mechanical ventilation. In a 2012 study by Alsadat and co-authors, the implementation of a VAP bundle is described as a performance improvement intervention rolled out over several months in the critical care units of central four hospitals and aiming to decrease VAP rates. The bundle project was implemented after several educational sessions on VAP were monitored using statistical process control charts. Because multiple factors contribute to the high risk of ventilator-associated pneumonia, a multi-strategy approach was developed through the ventilator bundle to prevent morbidity associated with ventilator use. The intervention included: 1. an initial educational workshop for the stakeholders from different participating hospitals with the emphasis on standards of practice; 2. nursing support and physician leadership; 3. the creation of a multidisciplinary team in every hospital; and 4. staff education with bedside mentoring and skills competency documentation for all the nurses and physicians in the specific unit. Teams conducted daily rounds on all ventilated patients and recorded compliance with elements of the VAP bundle. Data were collected daily and plotted on the Statistical Control Chart (SPC) on a weekly basis, then analysed at the end of each month to describe results and make recommendations. SPC charts were used to monitor the process of compliance with the individual bundle elements and the whole bundle. The study concluded that the VAP bundle had performed differently in different hospitals, suggesting that a multi-approach strategy was needed to reduce overall VAP rates.

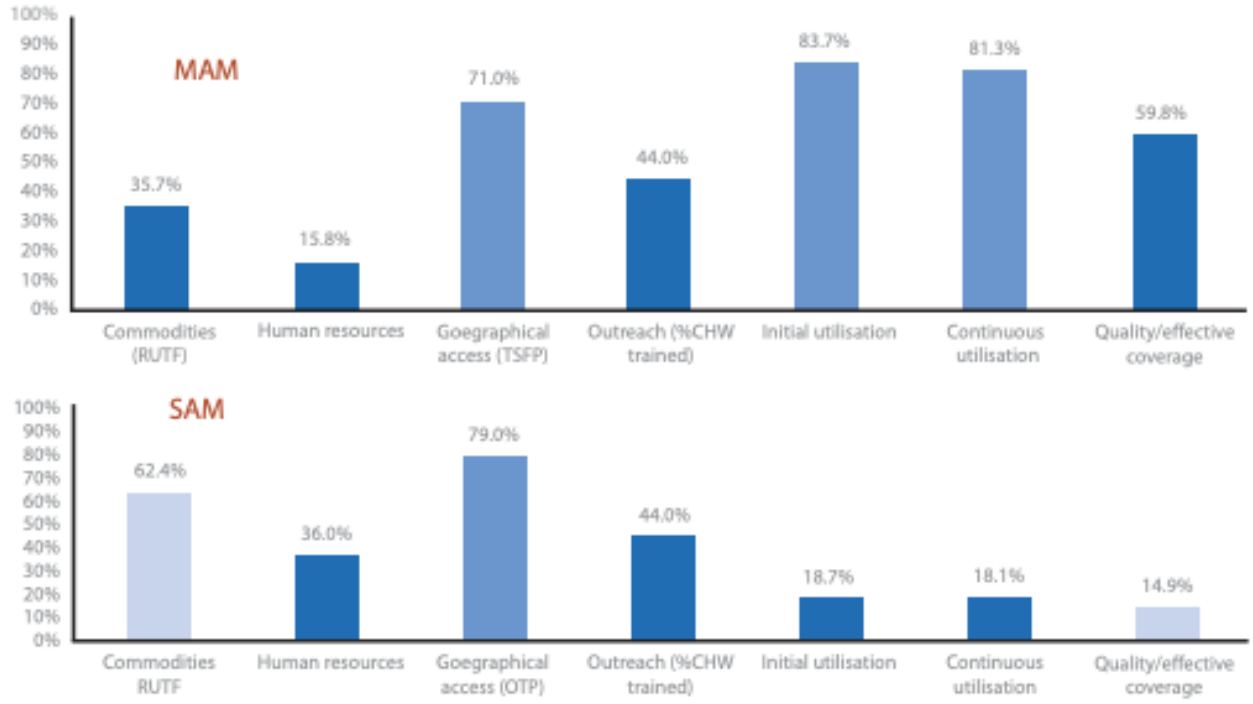
Speed: Enables to monitor performance over time and adjust evaluation process in real time.
Goal: Allows data users to draw logical decisions regarding the need for investigation and change in procedures.
Fit: Best applied throughout the life cycle of a project by means of time series plots that help assessing whether observed changes reflect random variation or 'real' change in the outcome of interest, with outcomes that are achieved quickly and at high frequency.

Bottleneck analysis (BNA) is a quantitative method to identify steps that link the intended beneficiaries of a programme or customers with actual beneficiaries. Each step is conditional on the previous one having been met, and only the population reaching the end of all the steps achieves the desired outcome. The relative size of the population lost at each step might indicate where the most urgent action is needed. BNA analysis can be stratified to understand differences between sub-groups. It is undertaken once an intervention is running and is often undertaken at a single point in time providing a snapshot of needs. Where routine or programme data are available,

BNA analysis can be repeated to assess whether the bottlenecks identified and their size change over time.

One application of this method in the international development landscape is a BNA conducted in Somalia by the country offices of UNICEF, the World Food programme, and the Department of Public Health as part of a comprehensive strategy to address the multi-dimensional causes of persistent malnutrition in the country. A bottleneck analysis was conducted to assess the determinants of effective coverage of the *integrated management of acute malnutrition* (IMAM) programme. The BNA process consisted of four distinct stages carried out over a 16-month period between June 2016 and October 2017 that showed sub-optimal coverage and integration of two segments of IMAM: treatments for children with 'moderate acute malnutrition' (MAM), and treatments for uncomplicated 'severe acute malnutrition' (SAM).

Figure 1: Somaliland IMAM BNA (1 September 2015 to 31 August 2016)



Source: Rio et al (2015)

The BNA was a consultative and participatory process to promote and build the capacity of the government and partners in the scale-up of the treatment of acute malnutrition. Following on the recommendation of the BNA, substantial progress has been made on the integration of MAM and SAM treatment, with most nutrition sites delivering both, proving the overall value of conduct BNA to develop a long-term human resource development strategy in humanitarian situations.

Speed: Uses monitoring data or existing data from the programme to identify barriers to project activities in real time.
Goal: Assess barriers and identify solutions to the implementation of an intervention.
Fit: Provides a rigorous framework for analysing factors determining access to the project through the identification of barriers (bottlenecks) and strategies to remove them within a specified time.

Contribution analysis is an impact evaluation approach designed to understand the contribution that a programme has made (or is currently making) to outcomes. The method uses six steps to refine a programme theory of change (ToC) through an iterative cycle, with particular attention given to contextual influences. Contribution analysis maps the ongoing activities that are expected to contribute to a particular outcome. It is a structured approach well suited to evaluating programmes that are regularly adapted to contextual changes, and that result from collaboration with multiple partners. The evidence for contribution analysis is collected through 'performance stories' (a short reports focused on the series of expected changes set out in the programme's theory of change) that are then used to confirm or revise the theory of change, provide feedback on what is driving change, and unpack the relative contribution of a particular activity to outcomes.

A recent ODI policy brief (Apgar et al, 2020) explains the use of contribution analysis in four international development programmes operating under conditions of complexity and uncertainty. One of the four programmes is the FCDO funded Child Labour: Action Research-Innovation in South and Southeastern Asia (CLARISSA, 2019-2023), which aims to reduce the worst forms of child labour in Bangladesh, Myanmar and Nepal. CLARISSA used several interlinked ToCs to map out the intervention at various levels: an abstract; general ToC; and more contextualised and specific ones for identifying and using actionable learning to adapt the programme. At the programme level, CLARISSA's ToC was co-developed with the participation of all consortium partners as well as the donor, whereas at the country level, the ToC was developed with the participation of at-risk children, implementation teams and other change agents in the child labour system using a participatory implementation modality. This initial ownership of the ToC enabled a richer evaluative process to identify emergent pathways and contribution claims to be critically assessed. As a result, CLARISSA was able to use contribution analysis to learn through the child-centred action research programme and to produce evidence and innovative solutions driven by people's definitions of their own problems.

Speed: Contribution analysis is carried out in parallel with an intervention to learn and adapt the theory of change as required.

Goal: To confirm or revise a programme's theory of change.
Fit: This is a structured approach well placed to evaluate programmes with overlapping activities that are regularly adapted to contextual changes.

Qualitative comparative analysis (QCA) is a theory-driven method that aims to identify the configurations of participants, intervention and contextual characteristics that may be associated with a given outcome. QCA starts with documenting different configurations of conditions associated with each observed outcome. The number of these configurations is then reduced to identify the simplest set of conditions that can account for all observed outcomes and their absence. QCA addresses multiple 'causation by transforming cases into configurations or combinations of factors that are referred to as 'conditions' that produce a given outcome of interest. The key question that QCA therefore seeks to address is which conditions (or combinations of conditions) are 'necessary' or 'sufficient' to produce the outcome. However, QCA establishes correlation between variables without implying causation. The causal character of the relationship between a configuration of conditions and an outcome is a theory-informed assumption, not a conclusion. The QCA approach is particularly useful to distinguish various complex forms of relevant causal factors that may affect outcomes. QCA makes even stronger assumptions than statistical methods that there no causally relevant factors that might contribute to a given outcome that have been omitted. In a standard regression analysis the error term can, in principle, capture the effects of omitted relevant causal variables. In contrast, in a variable-oriented approach, the solution to the problems posed by omitted variables is not to include as many potentially relevant variables in the model as possible explanatory ones, but rather through a careful research design that only includes variables in the analysis that are bound to be relevant, making the tests more focused and controlled. In this sense, the advantage of QCA is that it requires relatively small and simple data sets to achieve statistical significance. It is classified as a theory-driven approach because the choice of conditions being examined is informed by a priori theory on what matters. The coding of the presence/absence of a condition also requires an explicit rationale.

One recent application of QCA is an investigation of the migration–development nexus to explore under what conditions migration is a driver rather than an obstacle for development (in other words the two-way relationship between migration and development) (Czaika and Godin, 2021). QCA was used to examine which factors explain internet access provision and adoption in Ghana (Taylor, 2015). The paper discusses international mobility as a means for small-scale entrepreneurs to access technological resources and knowledge. It uses survey data gathered in 2009 from 95 internet cafés in Accra that were used to conduct QCA. Survey data included information on

respondents' past international movements and existing contacts, which suggested a large variation of mobility in terms of duration, destination and aims. QCA analysis identified two important findings. First, that international networks and accounting abilities function as substitutes for formal education (the latter two being unnecessary to run successful businesses. This finding provided grounds to reject the often-used argument that education is necessary for small business development. The second finding was the importance of international networking: those who migrated for work overseas began building lasting networks upon their return, and younger entrepreneurs with no start-up used online contacts instead of direct connections made while living abroad. The example illustrates how QCA is a method that allows the assessment of highly complex causal configurations, of conditions for explaining an outcome variable in often small-sized samples.

Speed: It is easy and fast to implement QCA based on relatively small datasets that supports short cycle learning about the effectiveness of specific activities being implemented during a project's lifespan.
Goal: To analyse the causal contribution of different aspects of an intervention to an outcome of interest.
Fit: When the evaluation comprises many comparable cases and facilitates the identification of necessary and sufficient conditions for an outcome to be obtained on the assumption that multiple pathways might lead to the same outcome.

3.3 Management methods

Problem-driven iterative adaptation (PDIA) is an approach first introduced in 2012 (Andrews et al., 2012) to help organisations in generating, testing, and refining context-specific solutions in response to locally identified problems. The method is centred on building capability through problem-solving, and therefore is not about finding the solution to a problem and replicating that solution elsewhere, but rather it is about understanding the problem-solving process. PDIA follows a step-by-step process that allows for flexible learning and adaptation strategies based on four core principles: problem-solving; allowing positive deviation; iterating and adapting; and scaling practices through diffusion. PDIA is most appropriately used for problems in the public sector that are simultaneously logistically complex, politically contentious (i.e., implementing them may generate potentially hostile resistance), and have no known solution prior to starting, but numerous opportunities for individual discretion to influence action. The method works through the iteration of six 'find and fit' stages that aim to support a gradual and progressive identification and implementation of adjustments needed for programme impact.

MUVA, an FCDO-funded programme (2015–2022) aims to help young women in urban areas of Mozambique to become economically empowered by creating opportunities for them to gain the education and skills (including soft skills) to acquire and retain secure, well-paid jobs, and have better access to markets. In 2017, women's access to formal employment in poor urban areas was found to be very weak, with most women having heavy childcare responsibilities and involved in small-scale trade by selling their produce with an average profit of around 1 USD dollar a day. MUVA used a PDIA approach (Hearle et al, 2019) to help women diagnose their own problems and identify the hindering factors to businesses' growth, allowing their views and experiences to shape the programme. More importantly, MUVA facilitated the development of problem-solving skills required for greater personal and professional growth amongst marginalised female market sellers.

Speed: Relatively rapid; locally identified and prioritised problems are solved as they arise by testing and refining context-specific solutions.
Goal: To build capability through problem solving.
Fit: Best fit for evaluations problems often carried out in the public sector that are logistically complex and politically contentious.

Rapid Cycle Design and Testing is a method to develop, design, test, refine and improve services. It is a flexible approach in evaluation that features a number of small, iterative tests, the results of which are used by organisations to improve their services. This is a programme improvement approach that provides programme leaders with evidence about what works to improve services. Programme administrators can use internal data already collected for monitoring purposes to generate evidence that rapidly informs how to improve service delivery. Researchers and service delivery organisations usually adopt and adapt rapid-cycle methods at the formative stage of service design to help understand the quality of their delivery, and whether their service, or parts of it, are making the difference they think it should.

One application of the method in international development is the rapid cycle evaluation of a community-based early intervention programme, the mobile Village Based Rehabilitation–Early Intervention (mVBR-EI). This programme provides access to early intervention (EI) services to children with delayed development in rural areas of Tamil Nadu, India, by leveraging mobile technology (Krishma et al, 2020). The primary objective of the programme was to increase access to early identification of disabilities and provide therapy services to children identified with disabilities to enhance children's physical, cognitive, communication, social and emotional development, and reduce caregiver burden. A rapid cycle evaluation brought almost immediate change to the programme. Qualitative and quantitative data were employed to inform the

process, with focus groups and interviews to generate information to inform programme changes, and longitudinal cohort study to observe potential effects of the changes on the outcomes sought by the programme. Based on the findings of the rapid evaluation, several rapid cycle actions were taken that led to greater programme engagement, improved school enrolment, and positioned the intervention for scale-up, also providing lessons that may be beneficial in other contexts.

Speed: Offers insights to programme funders, commissioners, providers, and users. It gives timely information for decision-making and improvement.
Goal: Monitor process data to provide rapid evidence of effective implementation in complex environments to learn and improve service.
Fit: Best used by organisations (researchers or implementers) wanting to consider external evaluation in the future but that wish to optimise the design and delivery of their service before the evaluation, or for organisations that have been externally evaluated and have identified weaknesses in design or delivery that they wish to address.

Developmental evaluation (DE) is an evaluation approach that can assist project/programme managers and staff develop social change initiatives in complex or uncertain environments. It is an approach, and not a method, for dealing with complexity in human systems and the need to provide structured, useful, actionable information to make decisions in real time. DE facilitates discussions around evaluative questions, and ongoing collection, analyse and use of information to support ongoing decision-making. Feedback can be provided on a continuous basis so that adjustments to projects and programmes can be made on an ongoing basis. DE is particularly appropriate for work in complex or uncertain environments, where the route to change is non-linear and cannot easily be predicted beforehand. DE has no fixed steps, and context determines the way it is applied. Systematic monitoring, formal or informal reviews, traditional evaluations, formal research, action-oriented research are all possible methods of application. Central to the value of DE is its use in the early stages of an intervention, when it can shape initial plans.

This approach was used in 2017 by USAID/Tanzania to assist the Government of Tanzania in better integrating its health services under Boresha Afya, a flagship health services project. CIRCLE was commissioned to conduct a DE over the life of the Boresha Afya Project until 2021. Using DE, the evaluation team was able to convene regional and national meetings to discuss emerging information about integrated services in three regions. Though a process of feedback loops, the Boresha Afya implementing partners identified over 160 actions and processes to improve service delivery. Actions included simple, impactful, and proven to be best practice that were, however, not happening.

Changes included ensuring that educational materials were available at clinics, identifying how to adapt clinic registers to reflect integrated services, communicating to district and regional health teams, pre-recording education sessions, and using appointment times to maximise the clinic day. DE was able to provide real-time learning to the project implementers and optimise health services based on their actual operating context.

Speed: DE facilitates real-time, or close to real-time, feedback to program staff, with the evaluation data gathered during DE being made sense of as the programme unfolds.
Goal: To help framing concepts, test quick iterations, track developments, emerging issues in the design and implementation phases of a complex interventions characterised by high degree of uncertainty.
Fit: Supports adaptation to complex or uncertain environments, where the route to change is non-linear and cannot easily be predicted beforehand.

3.4 Qualitative methods

Most significant change (MSC) is a form of participatory monitoring and evaluation qualitative method that features the collection and selection of stories of change developed from programme or project stakeholders. A panel of stakeholders selects what they consider to be the most significant stories and arrive at a reduced set of changes. MSC is normally used as an ongoing monitoring technique, assessing change throughout the lifetime of a programme or project. However, its focus on change (outcome and impact) means it can easily be adapted for use in evaluations as well. It is not a planning tool and is only normally used within a project or programme once enough time has elapsed for change to have occurred, so it is considered a retrospective method. MSC can help explain how change occurs (processes and causal mechanisms) and when (in what situations and contexts), and so it is useful in supporting the development of a programme's theory of change or logic models but also to assess the performance of the whole programme. Operationally, it works through three steps: 1. selection of the types of stories to collect; 2. collection of stories and assessment of the most significant ones; and 3. sharing the stories for discussion with stakeholders to enhance learning about the similarities and differences in what different groups and individuals' value.

A paper by Ho and co-authors (Ho et al, 2015) describes the implementation of community scorecards (CSC) within a community-driven reconstruction project called Tuungane (*Let's Unite* in Kiswahili) in two provinces of eastern Democratic Republic of Congo (DRC). The project aimed at understanding how the health system had been

affected after more than 10 years of conflict in the country. The MSC technique was used over an 18-month period in 2012/13 to collect stories of change in the health system from village development committee, health committee, community members and healthcare providers. The study focused on examining changes perceived as significant by the staff and the beneficiaries involved and showed some of the mechanisms by which CSC can improve the functioning of local health systems in conflict-affected settings by providing information to users and providers and encouraging them to engage in making health services more responsive to their needs. The findings suggested that in DRC — where the central government has limited influence on many aspects of what happens in the periphery — the divisions between frontline healthcare providers and community members can be bridged by facilitating space for interface, exchange, and collaboration. MSC programme staff saw collective action problems on both the supply and demand sides that needed to be overcome. The observations of how change is taking place suggested that a collaborative approach to accountability, with efforts from both users and health workers would help solve problems within local health services.

Speed: Approach used to assess change throughout the lifetime of a programme or project.
Goal: To learn about how and when intervention effects occur.
Fit: Useful method in contexts where the programme is characterised by a high degree of complexity when it is not possible to predict with any certainty what the outcome of the programme will be, when outcomes vary widely across beneficiaries, or when there is no agreement between stakeholders on which outcomes are the most important. It is also useful as part of a process evaluation, as complement to a quantitative impact evaluation.

Qualitative Impact Assessment Protocol (QuIP) is an impact evaluation approach developed by researchers at the University of Bath's Centre for Development Studies that provides a 'reality check' of a predetermined theory of change to support stakeholders in assessing the social impact of their work. It is an approach that puts the voice of intended project beneficiaries at the centre of the evaluation, to share their experiences in a credible way. QuIP gathers narrative causal statements directly from project beneficiaries to gather evidence on impact. The narrative statements are employed to ask respondents about the main changes in their lives over a pre-defined recall period, discuss drivers of change perceived to have contributed to these changes, and to whom or what they attribute any change – contemplating multiple attributes of change if required. The method works through key informants' attribution of the causal mechanism, and therefore does not rely on the identification of a control group

and no statistical inference is required. QuIP uses semi-structured interviews and focus groups conducted in the native language by highly skilled, local researchers blind to who commissioned the research or what project is being assessed to limit confirmation bias. Purposive sampling is done through a rigorous process until saturation. Where good monitoring data is available it can be used to decide the number, location and variation of respondents selected, based on differences in context, geography, treatment and/or positive and negative results from existing monitoring data. QuIP questionnaires are designed to reflect the areas of people's lives assumed to be affected from the project's theory of change, but the questions are framed around outcomes rather than inputs to collect information more broadly about what has changed. Once data are collected, QuIP systematically codes drivers of change, outcomes and reported attribution using a systematic and replicable approach to coding. The results generate key stories of change which show emerging trends and patterns between different respondent types.

A recent book by QuIP developers collects several case studies describing how the approach is applied in an international development context (Copestake et al, 2019). One of the case studies discusses the application of QuIP to Diageo, the largest global beverage company in the world who commissioned a study to conduct a 'deep dive' in its supply chain in Ethiopia. Diageo had consolidated in Ethiopia a niche market in 2012 by purchasing both local breweries near Addis Ababa and was interested to understand the wider (and possibly negative) social and economic consequences of its operations on the ground on smallholder malt barley farming, as well as finding insights into the impact of the programmes. The QuIP study confirmed positive income effects of Diageo's malt barley procurement in one area but highlighted unanticipated problems in another. It provided reassurance to the company that there was no issue of child labour involved with the activities. Findings were elaborated thanks to the availability and accessibility of good monitoring data to capture significant diversity in respondents' experiences.

Speed: QuIP is conducted throughout the lifecycle of an intervention and takes a systematic and replicable approach to coding which speeds the process of learning from beneficiaries' voices.
Goal: The QuIP aims to gather evidence about the causal processes of change, not to quantify impact.
Fit: Provides a reality check of a predetermined theory of change to support stakeholders to assess the social impact of their work. It is particularly useful in programmes where the context is highly changeable and helps inform if and how there is a need to adapt activities.

Rapid Rural Appraisal (RRA) is a methodology developed in the late 1970s in response to problems with long-delayed answers to large-scale, structured questionnaire surveys. RRA was designed to provide decision-makers with relevant, timely, accurate and usable information to learn quickly from 'insiders' (e.g., local community members and government officials) about their realities and challenges. RRA subsequently led to the development of Participatory Rural Appraisal, which focused more strongly on enabling the active participation of community members in generating knowledge about their own contexts and experiences through skilful facilitation and sustainable, locally driven rather than treating them as 'subjects' of international development research. RRA activities include three broad categories: 1. gathering background information from pre-existing data, and team discussion to develop preliminary hypotheses; 2. relatively short field visits (single or multiple) to the study areas; and 3. discussions and analysis among team members to reach consensus on what has been learned. The critical factor with RRA is that, by its sparing demands for information, it releases time for checking existing data, identifying unasked questions, and enabling more contact with, and learning from local people.

Dawoud and Hassan (2015) discuss the use of RRA as a descriptive participatory research tool to better understand how the Sudanese Darfur conflict widened rivalry between different ethnic groups, eroding the social fabric and their peaceful coexistence in most villages in Central and West Darfur. High incidences of localised insecurity along the entire corridor, concentration of returnees and IDPs (internally displaced persons) areas perceived safer had been causing human induced environmental degradation. The RRA was used through the design of questionnaires developed to capture livestock and the dynamics of livelihood trends across different groups in West and Central Darfur. Group discussion and key informants were organised, and data collected from various stakeholders (pastoralists, agro-pastoralists, sedentary herders and IDPs) in all localities in the region. The results were helpful to understand the rapid transition in livelihoods in the region. Traditional livelihood strategies linked to camel-based pastoralism have declined with the loss of access to seasonal pastures and the massive increase in salaried military service had become the major livelihood strategy. Pastoralist lifestyles had also changed as seasonal movements became restricted to safe zones, and movement restrictions denied them access to their favoured pastures. These changes in livelihoods were said to result from the impact of conflict.

Speed: This approach has low demand for information, it relies on oral and visual communication to ensure rapid capture of data, releasing time for other evaluation tasks.

Goal: To incorporate the knowledge and opinions of rural people in the planning and management of development projects and programmes, typically run by NGOs.
Fit: Adds important local knowledge in the planning, implementation, and evaluation of a project, providing greater understanding of the complex contextual dynamics.

Behaviour-Centred Design (BCD) is an approach which mixes both science (an evolutionary framework) and creativity to design successful behaviour change programmes (Aunger and Curtis, 2016). It is an approach to intervention design which can lead to rapid learning during the formative research phase of an intervention and includes some element of evaluation. The evaluation, however, focuses narrowly on behavioural impacts rather than impacts on further downstream health outcomes. The assumption underlying the approach is that behaviour will only change in response to something new and challenging. Successfully employed in a range of public health behaviour interventions as well as in commercial product design and marketing, BCD targets both individuals and societies. It starts by providing a behavioural model theory to develop a taxonomy of needs based on evolutionary biology. BCD then shows how the disruption of 'behaviour settings' happens, and finally identifies the steps involved in programming for behaviour change. BCD offers helps identifying the triggers that can change behaviour, and steps and tools to use in conceiving, creating, implementing, and evaluating a behaviour change programme.

An interesting application of BCD was made in Tanzania to inform the theory of change for a multi-channel National Sanitation Campaign (NSC) aimed at changing hygiene behaviour though empowerment and peer pressure, rather than through information about germs (Czerniewska et al, 2019). The adoption of a BCD approach resulted in a Theory of Change that recommended 'surprising' people with a new conversation about toilets, promoting toilets as a means of conferring status, and introducing the urgency to 'act now', and the notion that modest improvements would lead to a better life. The research provided considerable insight into sanitation behaviours in rural Tanzania, which in turn informed the design of innovative interventions.

Speed: Uses motivational mapping, product attribute ranking, and video ethnography to provide rapid 'deep dives' with target audiences that are then then used for formative research to describe a programme's Theory of Change.
Goal: To identify successful behaviour change programs through a design process, steps, and tools for conceiving, creating, implementing, and evaluating behaviour change.
Fit: Used in complex interventions hard to implement and evaluate where it helps focus on a very small number of behaviours and use a unified and easily identifiable intervention campaign to address the intended behavioural change.

3.5 Navigating through timely evaluations options

The methods reviewed above are not intended to span the entire spectrum of approaches to timely evaluation, as additional approaches certainly exist. We have focused on approaches that can be used in the context of international development programmes and complement more traditional experimental or quasi-experimental impact evaluations. The key feature of the approaches reviewed in this paper can be summarised as follows:

- Timely evaluations with **experimental design** approaches use statistical techniques to assess the relatively effects of different interventions.
- Timely evaluations using **monitoring design** approaches focus on learning how activities within an intervention are combined, or affect the changes expected. This enables evaluators to learn whether the intervention is unfolding according to its planned causal pathway and update it if necessary.
- Timely evaluations using **management** approaches are suitable to provide ongoing information on how programmes or services are delivered, and when and how they may require adaptation to changing circumstances. Their core feature is to provide flexible learning and adaptation strategies.
- Timely evaluations using **qualitative** approaches involve participatory methods that place project beneficiaries' and other stakeholders' views centre stage to test the ToC or develop the programme in real-time to maximise chance of effectiveness.

Many of these approaches have in common a desire to adjust the intervention while it is ongoing to maximise its potential effectiveness.

Table 2 also captures the key characteristics of each method reviewed based on the three criteria proposed in the analytical framework of the paper: speed, goal and fit to support readers in constructively thinking which methods are more apt to generate stronger evidence in impact evaluation research.

Table 2: Evaluation methods by family group, and 'timely' criteria

		Speed	*Goal*	*Fit*
Experimental	*Stepped wedges RCT design*	The method allows to understand important insights of how the intervention delivers its effect as exposure to treatment is sequenced across clusters.	The goal is to provide research insights in support of (or against) the intervention at the cluster level (though not at the policy level) as the intervention is rolled out.	Through the sequential transition of clusters from control to intervention conditions in a randomized order, SWTs are particularly useful in the evaluation of complex interventions that take considerable time to influence outcomes.
	A/B testing	Compares the reaction of users to different methods of intervening in real time.	Compares two or more versions of a content type to determine which one performs better.	Method used to run a large number of simultaneous independent tests to compare two versions of an intervention using large volumes of data that can be rapidly collected.
	Adaptive trial	Rapid adaptation within the timeframe of a trial.	Learn to adapt the design quickly based on interim data, evidence, and short-term impact.	When researchers/programme planners need to adjust an intervention while it is ongoing to maximise its potential effectiveness.
	Factorial design	Can be used as a rapid screen to identify factors that are of most importance in determining response to an intervention.	To discover which factors influence the outcome of the experiment and what levels of these factors lead to an experiment with the greatest sensitivity.	To look at different interventions either on their own or as combinations, to determine which works best.
Monitoring	*Statistical Control Process*	Enables to monitor performance over time and adjust evaluation process in real time.	Allows data users to draw logical decisions regarding the need for investigation and change in procedures.	Best applied throughout the life cycle of a project by means of time series plots that help assessing whether observed changes reflect random variation or 'real' change in the outcome of interest, with outcomes that are achieved quickly and at high frequency.
	Bottleneck analysis	Uses monitoring data or existing data from the programme to identify barriers to project activities in real time.	To assess barriers and identify solutions to the implementation of an intervention.	Provides a rigorous framework for analysing factors determining access to the project through the identification of barriers (bottlenecks) and strategies to remove them within a specified time.
	Contribution Analysis	Contribution analysis is carried out in parallel with an intervention to learn and adapt the theory of change as required.	To confirm or revise a programme's theory of change.	This is a structured approach well placed to evaluate programmes with overlapping activities that are regularly adapted to contextual changes.

		Speed	*Goal*	*Fit*
	Qualitative Comparative Analysis	It is easy and fast to implement QCA based on relatively small datasets that support short cycle learning about the effectiveness of specific activities being implemented during a project's lifespan.	To analyse the causal contribution of different aspects of an intervention to an outcome of interest.	When the evaluation comprises many comparable cases and facilitates the identification of necessary and sufficient conditions for an outcome to be obtained on the assumption that multiple pathways might lead to the same outcome.
Management	*Problem-driven iterative adaptation*	Relatively rapid; locally identified and prioritised problems are solved as they arise by testing and refining context-specific solutions.	To build capability through problem solving.	Best fit for evaluations problems often carried out in the public sector that are logistically complex and politically contentious.
	Rapid Cycle Design and Testing	Offers insights to programme funders, commissioners, providers, and users. It gives timely information for decision-making and improvement.	To monitor process data to provide rapid evidence of effective implementation in complex environments to learn and improve service.	For organisations wanting to consider external evaluation in the future but and wish to optimise the design and delivery of their service before the evaluation.
	Developmental evaluation	DE facilitates real-time, or close to real-time, feedback to program staff, with the evaluation data gathered during DE being made sense of as the programme unfolds.	To help framing concepts, test quick iterations, track developments, emerging issues in the design and implementation phases of a complex interventions characterised by high degree of uncertainty.	Best fit to supports adaptation to complex or uncertain environments, where the route to change is non-linear and cannot easily be predicted beforehand.
Qualitative	*Most significant change*	Approach used for assessing change throughout the lifetime of a programme or project.	To learn about how and when intervention effects occur.	Useful method in contexts where the programme is characterised by a high degree of complexity when it is not possible to predict with any certainty what the outcome of the programme will be.
	Qualitative Impact Assessment Protocol (QUiP)	QuIP is conducted throughout the lifecycle of an intervention and takes a systematic and replicable approach to coding which speeds the process of learning from beneficiaries' voices.	The QuIP aims to gather evidence about the causal processes of change, not to quantify impact.	It provides a reality check of a predetermined theory of change to support stakeholders to assess the social impact of their work. It is particularly useful in programmes where the context is highly changeable and helps inform if and how there is a need to adapt activities.
	Rapid Rural Appraisal	This approach has low demand for information; it relies on oral and visual communication to ensure	To incorporate the knowledge and opinions of rural people in the planning and management of development	Adds important local knowledge in the planning, implementation, and evaluation of a project,

		Speed	*Goal*	*Fit*
		rapid capture of data, releasng time for other evaluation tasks.	projects and programmes, typically run by NGOs.	providing greater understanding of the complex contextual dynamics.
	Behaviour Centred Design	Uses motivational mapping, product attribute ranking, and video ethnography to provide rapid 'deep dives' with target audiences that are then then used for formative research to describe a programme's Theory of Change.	To identify successful behaviour, change programmes through a design process, steps, and tools for conceiving, creating, implementing, and evaluating behaviour change.	Used in complex interventions hard to implement and evaluate where it helps focus on a very small number of behaviours and use a unified and easily identifiable intervention campaign to address the intended behavioural change.

4. Discussion and implications for CEDIL research agenda

This paper set out to develop a framework to categorise approaches to timely evaluation. We identified both quantitative and qualitative approaches with different purposes, suggesting that selection of any approach should be made based on the specific time-needs and flexibility of the programme, noting that a mixed-methods approach generally maximises evaluation learning.

Detecting change in a timely manner relies on the analysis of outputs and short-term outcomes to indicate change rather than longer-term impacts. This particularly applies to quantitative methods such as SPC, A/B testing and interim analysis of adaptive or modified trials. The use of shorter-term outcomes runs the risk of falsely detecting treatment effects at a time in the intervention when there are many moving parts (changes in context, population behaviour affected by short term response to programme activities) or prematurely discarding promising interventions that do not show an impact at an early stage. It is therefore important to recognise that the short-time horizon of applicability of the findings and conclusions drawn need to be viewed with caution as assessing impact over a longer period might lead to different conclusions or other information emerging as causal processes work over different time scales (Woolcock, 2009).

These methods should be combined with qualitative methods to pick up unanticipated outcomes. When using methods such as SPC, that have the flexibility to change the outcomes measured over time, researchers should consider the value of including some constant or 'bedrock' indicators that don't change over the life of the programme to support an understanding of the longer-term impact of projects (Barr, 2015).

Timely evaluation approaches are likely to be more resource intensive and require more data to be collected than traditional evaluation methods. Where a specific causal mechanism is not being tested, evaluators will need to capture a wider range of outcomes. Methods that aim to rapidly test changes or compare multiple interventions rely on ongoing or repeated collection of data. The methods are nevertheless anticipated to represent overall value for money as they are expected to result in the programme having a higher chance of success.

Timeliness obviously depends on the ability to collect, process, and analyse data in a timely fashion. A key challenge is to better leverage data from service delivery platforms and to make such data useful (i.e., capture relevant outcome indicators in a timely manner) and of sufficient quality (i.e., measures needed to enhance

completeness and accuracy of data). One potential solution is to extended monitoring data, which typically captures data on inputs, activities and outputs, to include outcome indicators; in some of the literature this is described as 'impact monitoring' (Kessler and Tanburn, 2014).

Participants at the initial events organised in preparation for this paper observed that technology offered great potential to support more timely data collection, even in very resource-poor and challenging settings (DFID, 2012). For example, the American Refugee Committee uses digital technology to collect highly focussed satisfaction data from refugees in camps in Uganda, Rwanda, Somalia and Sudan (Peters, 2018). During the Ebola outbreak in West Africa 2013-2016 real-time surveys were undertaken by a number of partners (Cori et al., 2017). Similarly, the analysis of big data is already commonplace in the private sector, with consumer profiling and predictive analysis used extensively for advertising and service improvement (UN Global Pulse, 2012). Technology that offers increasing opportunities for real-time data analytics and their application should be explored more in development programmes.

Uncertainty over what evidence might be needed and when, is often compounded by delays in the time it takes commissioners and evaluators to respond. Moreover, timely evaluations require implementers and evaluators to work together to determine realistic timeframes within which changes can be expected, data can be collected, analysed, and decisions taken based on feedback. Stakeholder engagement can increase the utility and uptake of evaluation results – an approach termed 'utilization-focused evaluation' (Patton, 2008) - in which end-users are engaged from the outset to guide decisions about the evaluation process.

The range of timely evaluation methods presented in this paper can be employed over varying timeframes and to answer different evaluation questions. We have suggested a new categorisation to guide evaluators on what methods are best placed to address evaluations design problems, and implementation problems.

The methods reviewed should mainly be used to refine the initial design of an intervention or to decide whether such an intervention is meaningful (i.e., whether there will be enough uptake to warrant moving forward). However, each method has its limitations (for example, stepped wedge trials carry a potential bias from temporal trends, as does the statistical process control method).

It is necessary to recognise that these methods are not able to resolve the identification problem that enables to measure the causal impact of an intervention, certainly not in the long run. For this reason, it is important to consider that the combination of impact evaluation and timely evaluation methods for adaptive learning

is the best approach for a full rounded understanding of what works, when, and for whom.

How is CEDIL approaching the issue of timeliness in impact evaluation research? Rapid evaluations are becoming more frequent in international development contexts. Evaluators need to be responsive to changing priorities and deliver evaluation findings within shorter time frames to ensure that policy innovations can be evaluated rapidly to inform decisions about ongoing funding and scale-up. CEDIL's research agenda is rapidly adapting to these needs, with programmes of work commissioned to address complex evaluations, the use of big data and machine learning methods to maximise the benefits of data, and analytical tools to aid rapid decision-making based on existing information. CEDIL remains committed to resolving the ongoing challenge of delivering research and evaluations that align with the three Rs discussed by Riley and co-authors (2013): rapid, responsive, and relevant. We hope that this paper will help programme planners and evaluators to expand the portfolio of timely evaluation approaches at their disposal.

References

Alsadat, R., Al-Bardan, H., Mazloum, M. N., Shamah, A. A., Eltayeb, M. F., Marie, A., Dakkak, A., Naes, O., Esber, F., Betelmal, I., & Kherallah, M. (2012). Use of ventilator associated pneumonia bundle and statistical process control chart to decrease VAP rate in Syria. *Avicenna journal of medicine*, *2*(4), 79–83. https://doi.org/10.4103/2231-0770.110736

Andrews, M., Pritchett, L. & Woolcock, M. (2012) *Escaping Capability Traps through Problem-Driven Iterative Adaptation (PDIA)*. Working Paper 299. Centre for Global Development. Available at: https://www.cgdev.org/publication/escaping-capability-traps-through-problem-driven-iterative-adaptation-pdia-working-paper (accessed 11 October 2017).

Apgar, M., Hernandez K., & Ton G. (2020) Contribution analysis for adaptive management. ODI Briefing note (last consulted 13 March 2021 on: https://www.odi.org/sites/odi.org.uk/files/resource-documents/glam_contribution_analysis_final_1.pdf

Aunger, R. & Curtis, C. (2016) *Behaviour Centred Design: towards an applied science of behaviour change*, Health Psychology Review, 10:4, 425-446, DOI: 10.1080/17437199.2016.1219673

Barr, J. (2015) Monitoring and Evaluating Flexible and Adaptive Programming. In: *Itad*. Available at: http://www.itad.com/monitoring-and-evaluating-flexible-and-adaptive-programming/ (accessed 2 November 2017).

Burke, H.M., Chen, M., Brown, A.N. (2018) South Africa HIV and AIDSHIV preventionIntegrated development technical services Economic developmentHealth *IDS Bull 2018 Sep; 49(4): 115-28.*

Caria, S., Gordon, G., Kasy, M., Quinn, S., Shami, S. & Teytelboym, A. (2020), "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan", CESifo Working Paper Series 8535, CESifo.Cellamare M, Ventz S, Baudin E, et al. (2017) A Bayesian response-adaptive trial in tuberculosis: The endTB trial. *Clinical Trials (London, England)* 14(1): 17–28. DOI: 10.1177/1740774516665090.

Copestake, J., Morsink, M. & Remnant, F. (eds) (2019) *Attributing Development Impact: The QuIP Case Book*. Rugby: Practical Action.

Copestake, J., Allan, C., Bekkum, W. van, et al. (2018) Managing relationships in qualitative impact evaluation of international development: QuIP choreography as a case study. *Evaluation* 24(2): 169–184. DOI: 10.1177/1356389018763243.

Cori, A., Donnelly, C.A., Dorigatti, I., et al. (2017) Key data for outbreak evaluation: building on the Ebola experience. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372(1721). DOI: 10.1098/rstb.2016.0371.

Czerniewska, A., Muangi, W.C., Aunger, R., Massa, K. & Curtis, V. (2019) Theory-driven formative research to inform the design of a national sanitation campaign in Tanzania. PLoS ONE 14(8): e0221445. https://doi.org/10.1371/journal.pone.0221445

Czaika, M. & Godin, M. (2021) Disentangling the migration-development nexus using QCA, Migration and Development, DOI: 10.1080/21632324.2020.1866878

Dawoud, A. A. & Hassan, T. A. (2015) The Dynamics of Pastoralist Livelihoods in Central and West Darfur States as a Result of Darfur Conflict. J*ournal of Agricultural Science and Engineering* Vol. 1, No. 2, pp. 83-88

DFID (2012) *Results in Fragile and Conflict-Affected Stated and Situations: How to Note*. 28 February. Department for International Development. Available at: https://www.gov.uk/government/publications/results-in-fragile-and-conflict-affected-states-and-situations.

Hargreaves, J.R., Copas, A.J., Beard, E. *et al.* Five questions to consider before conducting a stepped wedge trial. *Trials* 16, 350 (2015). https://doi.org/10.1186/s13063-015-0841-8

Hearle, C., Baden S., & Kalsi, K. (2019) Promoting economic empowerment for women in the informal economy. *WOW Helpdesk Guidance No. 1 (accessed on March 1st, 2021 at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/844320/Promoting-Economic-Empowerment-Women-Informal-Economy-Oct19.pdf)*

Ho, L.S., Labrecque, G., Batonon, I., et al. (2015) Effects of a community scorecard on improving the local health system in Eastern Democratic Republic of Congo: qualitative evidence using the most significant change technique. *Conflict and Health* 9: 27. DOI: 10.1186/s13031-015-0055-4.

Kessler, A. & Tanburn, J. (2014) *Why Evaluations Fail: The Importance of Good Monitoring*. The Donor Committee for Enterprise Development.

Krishna, D., Muthukaruppan, S.S., Bharathwaj, A., Ponnusamy, R., Poomariappan, B.M., Mariappan, S., Beevi, A., MacLachlan, J., Campbell, Z., Anthonypillai, C., Brien, M., Cameron, C., Flatman, M., Perlman, L., Seilman, S., Jeyapragash, A., van der Haar, L., Krapels, J. & Srinivasan, S.R. (2020) Rapid-Cycle Evaluation in an Early Intervention Program for Children With Developmental Disabilities in South India: Optimizing Service Providers' Quality of Work-Life, Family Program Engagement, and School Enrolment. *Frontiers in Public Health* 8:567907. doi: 10.3389/fpubh.2020.567907

Patton, M.Q. (2008) *Utilization-Focused Evaluation*. Sage publications.

Peters, A. (2018) At these camps, refugees can give real-time customer feedback. Available at: https://www.fastcompany.com/40575160/at-these-camps-refugees-can-give-real-time-customer-feedback (accessed 24 July 2018).

Riley, W., Glasgow, R. E., Etheredge, L. & Abernethy, A. P. (2013). Rapid, responsive, relevant (R3) research: A call for a rapid learning health research enterprise. *Clinical and Translational Science*, 2, 10.

Rio, D., Hedges, J., Woodhead, S., et al. (2015) *What is the bottleneck analysis approach for the management of severe acute malnutrition?* UNICEF and Action Against Hunger. Available at: http://www.coverage-monitoring.org/wp-content/uploads/2015/12/BAA-12-08-2015.pdf (accessed 21 January 2018).

Taylor, L. (2015). Inside the black box of internet adoption: The role of migration and networking in internet penetration in West Africa. *Policy & Internet*, 7(4), 423–446.

The Gambia Hepatitis Study Group. The Gambia hepatitis intervention study. Cancer Res 1987;47:5782-7

UN Global Pulse (2012) *Big Data for Development: Challenges and Opportunities*. United Nations. Available at: https://www.unglobalpulse.org/projects/BigDataforDevelopment (accessed 26 October 2018).

USAID Boresha Afya Developmental Evaluation. Available at: https://www.harpnet.org/project/boresha-afya-developmental-evaluation/

Vowles, P. (n.d.) Adaptive programming. Available at: https://dfid.blog.gov.uk/2013/10/21/adaptive-programming/ (accessed 11 October 2017).ID42

Wohlin, C. (2014) Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, New York, NY, USA, 2014, pp. 38:1–38:10. EASE '14. ACM. DOI: 10.1145/2601248.2601268.

Women World Banking (2018) *It's as Easy as A-B-C! How A/B Testing Helped Onboard More Women Customers through Mobile Account Referrals* accessed on March 2ns 2021 from: https://www.womensworldbanking.org/insights-and-impact/its-as-easy-as-a-b-c-how-a-b-testing-helped-onboard-more-women-customers-through-mobile-account-referrals/?utm_source=ideas42+Staff&utm_campaign=9f5b8d1b3d-EMAIL_CAMPAIGN_2018_06_08_03_01&utm_medium=email&utm_term=0_c304491931-9f5b8d1b3d-186138741

Woolcock, M. (2009) Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy. *Journal of development effectiveness* 1(1): 1–14.

Annex A Initial work to define the research agenda: workshop and public symposium: June 2017 – November 2017

Table 3: List of speakers and presentations at the scoping seminar

Speaker	Presentation
Joana Busza	Real time evaluation on harm reduction and harm mitigation for young people that migrate to Gulf State/Middle East for work
Tanya Marchant	IDEAS phase 2; Gombe State, Nigeria
James Hargreaves	CeSHHAR: The sisters with a voice programme
Lucy Reynolds	Reflections on practical experience of undertaking evaluations in different contexts
Emily Balls	Adaptive management and learning through outcome mapping
Paul Mee	Real time vs. required time. Optimising data cycle times in the monitoring and evaluation of programmes

Table 4: List of speakers and presentations at the public symposium

Speaker	Organisation	Presentation
Liz Allen	LSHTM	Statistical Process Control
Annette Boaz	St George's	Conducting timely and useable evaluation for programme improvement: Is stakeholder engagement the answer?
Jean Boulton	University of Bath	Embracing Complexity implications for programme design and evaluation
James Copestake	University of Bath	Relationships, technique, skill and (yes) timing: the choreography of credible and useful impact evaluation using the QuIP.
Val Curtis	LSHTM	Using Behaviour Centred Design for timely intervention design
Martin Dale	PSI	Bridging the gap between data and the implementation narrative using DHIS2
Clair Hutchings	Oxfam	Timely Evaluation – but whose evaluation is it?
Aly Visram	OPM	Reflections on 'timely evaluation for programme improvement'
Jayne Webster	LSHTM	Timely evaluation for programme improvement

Annex B Summary of group breakout session

Breakout group participants were asked to summarise their discussion in a tweet (see Table 5). There was considerable overlap in the discussion between groups and we note a couple of the key themes:

Understanding casual pathways was deemed to be essential for interpreting change. There is a need to articulate casual pathways from an early stage to be able to recognise when things are not going as planned.

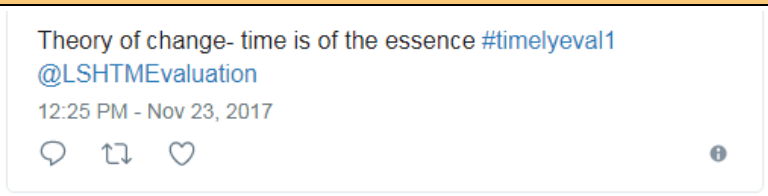
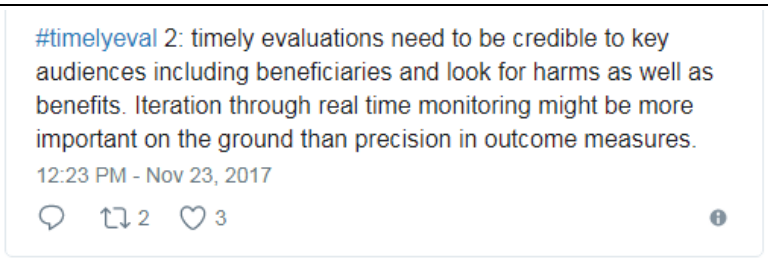
Evaluation findings need to be aligned with a programme's natural time cycles to be able to inform decision making.

Stakeholders' perspectives and engagement are key.

Whilst participants considered that a timely evaluation should be no different from an impact or process evaluation, there was acknowledgment that "timely" may result in a trade-off between timeliness and rigor, with, in some cases, the former being more important. Where speed is the key criteria, evaluators need to think about the scope of what is feasible and select methods accordingly.

Participants highlighted the need to learn from failure.

Table 5: Summary of group breakout session discussion

Question	Tweet
Theory of Change: How do we incorporate time into a Theory of Change? How do we estimate time for outputs to achieve outcomes? Can the Theory of change be an explicit approach to timely evaluation?	
What should be the major considerations when selecting methods for doing timely evaluation? How important are considerations of rigour, inference, biases etc.?	

Question	Tweet
What are the potential approaches? And what makes them timely?	Capture informal conversations to fasttrack data collection, whilst balancing robustness of methodologies, inc. analysis, with time available for, and scope of the evaluation. #timelyeval 3 12:26 PM - Nov 23, 2017 3
Do we need causal inference in doing timely evaluation for programme improvement?	.@LSHTMevaluation #timelyeval 4 Understanding causal & implementation pathways is essential for interpreting changes over time in programmes & is therefore necessary when doing timely evaluations for programme improvement 12:16 PM - Nov 23, 2017 1
How should approaches to impact and process evaluation differ when assessing timely evaluation programmes?	#timelyeval #gp5 ideally no differences it should be a continuum process but context frequency speed and costs can hinder the rigour/ different sectors mean different requirements for timeliness taking into account natural time cycles matter 12:26 PM - Nov 23, 2017 1 1
Have we got a research system that is fit for our desire for timely evaluations?	#timelyeval 6 Not yet, issues with unaligned stakeholder incentives, process and evaluation silos is a problem recognised and complex intervention methodologies are moving in the right direction 12:21 PM - Nov 23, 2017 1 3
Does 'timeliness' mean less account taken of community views?	If timeliness equals quick, restricts scope. Consider who is listened to, what's said, side effects, timing trumping reality #timelyeval 7 @LSHTMEvaluation 12:23 PM - Nov 23, 2017 2
What makes a programme 'adaptable'?	#timelyeval 8 on adaptive programming: "need to build in focus on ends not means, with a mindset to learn from failure, creatively" 12:23 PM - Nov 23, 2017 2 6
Does 'timeliness' come at the expense of quality and cost? What's the balance?	Appropriate evaluation tools used at right time can improve quality and reduce cost of programmes. The Right tools are different if eval/monitoring early, during or post. Timeliness, cost and quality all require inclusive adaptive management...? #timelyeval 9 12:17 PM - Nov 23, 2017 1

Question	Tweet
Is the concept of adaptive management a myth? Can it really happen in programmes implemented at scale? What are the implications for donors, evaluators, implementers?	#timelyeval 10 Adaptive management requires legitimising and learning from failure and buyin from funders, implementers and evaluators. Early reflection on process and knowing when to change paths is essential 12:26 PM - Nov 23, 2017 3 7

Annex C Search terms

Table 6: Search terms

Search	Terms (title/abstract/key word)	Hits	AND Exclusion criteria	Hits
1	"adaptive learn*" OR "continuous evaluat*" OR "developmental evaluat*" OR "experiential learn*" OR "feedback" OR "formative evaluat*" OR "real time evaluat" OR "Problem Driven Iterative Adaptation"	564,602	Document type: review or article	367,676
2	Humanitarian OR International Development	27,578	Document type: review or article	20,044
3	1 AND 2	200		129

Centre of Excellence for Development Impact and Learning

www.cedilprogramme.org

@CEDILProgramme

Centre of Excellence for Development Impact