Timely evaluation in international development

Jayne Webster[a*], Josephine Exley[b], James Copestake[c], Rick Davies[d] and James Hargreaves[b]

[a] *Centre for Evaluation and Disease Control Department, London School of Hygiene and Tropical Medicine (LSHTM);* [b] *Centre for Evaluation and Department of Social and Environmental Health Research, LSHTM;* [c]*Centre for Development Impact, University of Bath;* [d]*Independent Consultant, UK*

*corresponding author Jayne.Webster@lshtm.ac.uk London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

Timely evaluation in international development

Abstract:

Impact and process evaluations are increasingly used in international development, however they are generally retrospective in outlook. A more timely approach to evaluation aims to identify necessary, feasible and effective changes during a programme or intervention's lifetime. This paper aims to identify, categorise, describe and critically appraise methods to support more timely evaluation in international development.

Potential methods were identified through scoping seminar, public symposium, targeted review of the literature, and the authors' own experiences and opinions. Findings from the different data sources were reviewed collectively by the author group and triangulated to develop an analytical framework.

We identified four purposes of timely evaluation for international development, and critiqued the use of these approaches against four dimensions of timeliness and flexibility. Whilst we found significant interest in more timely approaches to evaluation in international development, there was a dearth of published empirical evidence upon which to base strong recommendations.

There is significant potential for timely evaluation to improve international development outcomes. New approaches to mixing and adapting existing methods, together with new technologies offer increased potential. Research is needed to provide an empirical evidence base upon which to further develop the application, across sectors and contexts, of timely evaluation in international development.

Keywords: outcome evaluation, impact evaluation, adaptive learning, programme improvement

Introduction

Outcome evaluations assess the impact of a specified set of actions, constituting a programme or intervention, on its intended outcomes. Such evaluations ask: what effect did this action have on these outcomes (often in comparison with some other action). Process evaluations

seek to explain how and why such impacts, did or did not, come about (Moore et al., 2015). They assess how implementation of a programme happened, whether hypothesised causal pathways were activated and identify contextual factors that acted as barriers or facilitators to either implementation, effectiveness, or both. Such evaluations are essential for informing future policy decisions, but many of the questions typically addressed are, by their nature, retrospective in outlook.

Dealing with the uncertainty and complexity inherent in international development settings requires a flexible approach to the design and implementation of programmes. Flexibility is needed across time (for example, changing activities or shifting priorities over time) and space (for example, adapting an approach to different settings and contexts), and happens at multiple speeds (for example, daily fine tuning of specific activities, annual changes in budget allocations and longer-term priority setting) (Barder and Ramalingam, 2012; Gamble, 2006; Ladner, 2015; Valters et al., 2016; Walji and Vein, 2013). The *Doing Development Differently* manifesto highlights that, among other things, to be successful development programmes need to 'merge design and implementation' by undertaking 'rapid cycles of planning, action, reflection and revision' and 'manage risk by making small bets; pursuing activities with promise and dropping others' (DDD, 2014).

Evaluations have a role to play in supporting the Doing Development Differently agenda by generating evidence to inform action during a programme's life cycle; from design to the selection, refinement and testing of interventions. Where knowledge is high about what is likely to work, evaluation can test whether the intervention is having the anticipated effect and support, and test, modifications over time. Where it is less clear what intervention might work, interventions need to be developed and options tested either sequentially or in parallel (Green, 2015; O'Donnell, 2016).

Despite there being a number of existing approaches and methods to incorporating evidence based decision making into programmes, there has been scant focus on, or critique of, 'timeliness' and the suitability of evaluation methods within flexible or adaptive international development programmes. We aim to review and critically appraise evaluation methods to support a more 'timely' approach to evaluations of international development programmes. To support this critical appraisal we define a 'timely' approach to evaluation and consider purposes of the evaluation and dimensions of the methods required for timely application and decision-making. To guide evaluators we propose a framework to support the selection of methods, or mixes of methods, needed to address particular evaluation questions at different stages of a programme's cycle.

Methods

Our review and critique of methods for timely evaluation included: a scoping seminar and public symposium to identify methods from the perspectives of academics, programme designers and programme evaluators; a review of approaches and methods used to evaluate international development programmes; and a critique of methods against a timely evaluation framework.

Scoping seminar and public symposium

The scoping seminar on 'real time evaluations for programme improvement' took place in June 2017 at the London School of Hygiene and Tropical Medicine (LSHTM) to harness the ideas and experiences of members of LSHTM's Centre for Evaluation. The seminar was attended by approximately 30 members from a range of disciplines within public health. The seminar included six speed talks and a group discussion. The public symposium held in November 2017 was attended by 142 people and included three sessions on: doing, evaluating and critiquing timely evaluations for programme improvement. Presentations were

given by eight speakers. During the event we engaged with participants through breakout sessions and technology. We had an active twitter discussion (#timelyeval) and used slido.com for participants to submit questions/comments during presentations. Both events were recorded and in drafting this manuscript we listened back to the recordings and took notes. Through the presentations and group discussions at the two events we collated a list of potential methods to examine in more detail.

Review of approaches and methods to evaluate international development programmes

The literature review consisted of two components. First, following the scoping event, we undertook a targeted review using a snowballing technique to identify specific methods that have been used in evaluations of adaptive learning approaches in development settings (Wohlin, 2014). Based on the scoping seminar, we developed a set of search terms (Table 1). Searches were run in PubMed and Web of Science. The reference list of relevant literature was screened, and we undertook forward citation searching in Google scholar. Second for the specific methods identified during the two events, targeted searches were run in google, google scholar, PubMed and Web of Science to identify examples of where the methods had been used in international development contexts.

Critique of methods against the timely evaluation framework

We developed a framework for timely evaluation of international development programmes and interventions based on our interpretation of the discussion at the scoping event, public symposium, and review of the literature. We critiqued examples of the methods against the timely evaluation framework.

Results

Based on the discussions at the scoping event and public symposium we defined a timely approach to evaluation as ‘the use of evaluation methods before or during the course of an

international development programme or intervention to provide evidence for decision making on design, adaptation or refinement at a time when these changes can plausibly lead to the improvements needed, and when implementers and stakeholders can effectively carryout and benefit from the changes'. This definition highlights the interconnected nature of timeliness and flexibility, which we expand on below.

During the internal and external events participants highlighted an array of existing approaches that they considered encapsulated aspects of a timely approach to evaluation including programme cycles, quality improvement, rapid cycle evaluations and developmental evaluations. Additional related approaches were identified through the literature review. At their core these approaches aim to generate more timely evidence over a programme or interventions life cycle and respond to changing and evolving priorities. The complete list of approaches identified are listed in Table 2.

The approaches listed in Table 2 often consist of a number of different methods. The challenge for evaluators is to identify suitable methods that can be used over varying timeframes to answer different evaluation questions at different time points as the programme unfolds. We summarise the methods identified through the scoping seminar, symposium and literature review in Table 3. The methods are both quantitative and qualitative, retrospective and prospective in their outlook, involve differing levels of technical skills in their analysis, and are generally applied at different stages of and time points within programmes for different purposes.

Framework for timely approach to evaluation

To support the selection of methods we conceptualise a timely approach to evaluation around an analytical framework (Figure 1). The framework consists of four overarching purposes and four timeliness and flexibility dimensions. The framework recognises that methods can

be used at different time points in the programme cycle and that the methods have different levels of flexibility that will make them more or less suitable in specific settings and contexts.

Purpose

The overarching purposes identified are: support design; identify problems; test potential solutions; and explain the outcomes.

Support design: of an intervention or package of interventions within a programme, conducted prior to and/or during implementation. Where data are collected prior to implementation the purpose is to make suggestions about what interventions should be implemented and how; or to determine modifications needed to a pre-existing intervention to implement in a new context. Where a programme or intervention is already running the purpose is to explore why an anticipated change might not have occurred and identify new interventions, changes to intervention design, or implementation strategies for existing interventions in reaction to identified problems.

Identify problems: where an intervention or programme is running the purpose is to monitor the status of implementation and identify problems that might need to be responded to. Monitoring may include all or a selection of components of a programme. Achievements are assessed against expectations which may be defined pre- or during implementation.

Test potential solutions: where need has been identified, the purpose is to test potential options and explain why they do or do not succeed in achieving the changes required. That is, evaluating whether particular interventions or course corrections are successful in meeting their stated objectives, or are comparatively better than other options, at a given time point during the programme.

Explain the outcomes: where problems in implementation or achievements have been identified and options/solutions are tested, it is important to understand and explain the

outcomes. Understanding how the tested solutions change the interventions, programmes or their implementation to facilitate improvement and increase the potential for learning.

The four purposes are not anticipated to proceed in a cyclical manner. For example, where a new design is identified or modification made the next step may be to test potential solutions or where a problem is identified then further research may seek to support the design of potential solutions to the problem.

Timeliness and flexibility dimensions

We identify four timeliness and flexibility dimensions that can be used to select between methods for specific purposes: design; speed; capacity; and space. The choice of method will depend on the required level of flexibility and potential time constraints. The dimensions should be considered together as they are overlapping and exert mutual influences one to the other.

Design: the extent to which a method can respond to emerging insights and unexpected or unintended consequences once it has been designed, gained approvals, and its implementation is underway.

Speed: ability of the method to adapt to time constraints and requirements. It considers the time required for design, data collection, analysis, reporting and feedback of data, and the potential to speed the process up.

Capacity: the level of skill required for design, data collection and analysis, and the extent to which there is flexibility around any of these.

Space: the ability of the method to adapt to different places and contexts.

Critique of methods against framework for application in international development

To illustrate the use of the analytical framework we mapped a sub-set of methods against the four purposes and critiqued the applicability of the methods for a more timely approach using

the four dimensions of timeliness and flexibility (Table 4). It is likely that over the course of a programme or intervention different methods will be needed to answer different evaluation questions and that the timescales and context will place restrictions on the suitability of different methods. A number of the methods identified can be used for multiple purposes and in general are not stand alone. We discuss the application of these methods for different purposes and discuss some of the challenges identified in critiquing the methods against the dimensions. The methods selected are intended to provide examples of the use of the framework to determine the applicability of a method, they are not intended to indicate exclusivity of these particular methods for timely evaluation in international development.

Support design

Both qualitative and quantitative methods can be used to support the development and/or refinement of an intervention or programme. Examples include rapid assessment process (RAP), a method of highly focussed ethnographic research, which draws on qualitative methods including in-depth interviews (IDIs), focus group discussions (FGDs) and observations (Beebe, 2001), and A/B testing (also known as nimble RCTs, split tests, rapid-fire tests, bucket testing, randomized field experiments), a randomised trial in which participants are randomly assigned to receive a variation of the same intervention (Dibner-Dunlap and Rathore, 2016; IPA, 2016; Karlan, 2017).

RAP is undertaken at a single point during the study to quickly develop a preliminary understanding of a situation. RAP was initially developed to support the evaluation of farming systems within a single planting season (Butler, 1995; Hildebrand, 1981) and has been used to develop interventions in health for example, to inform the development of tailored interventions for oral rehydration salts for diarrhoeal disease prevention within a limited time (Manderson and Aaby, 1992) and for assessing operational challenges in the

delivery of Long Lasting Insecticidal Nets (Theiss-Nyland et al., 2017). Qualitative methods such as IDIs and FGDs are able to adapt to rapidly changing contexts or shifting priorities over time; inductive adaptation of interview guides and discussion themes on a daily basis can respond to emerging or unexpected findings. Transcription, translation, coding and analysis for in-depth exploration of the data are time consuming but in RAP for example, adaptations for rapid use are made that enable completion of a study within a relatively short time period. Teams of interviewers may be used to rapidly collect information with the study completion expected within four to six weeks (Harris et al., 1997; Vlassoff and Tanner, 1992). The emphasis is on adequacy of data for the purpose, rather than high level of precision.

RAP methods can be undertaken before programme implementation, when there is ambiguity about the scale and nature of the problem and what is needed to address a problem. It can be used to characterise the setting, assess whether a proposed programme or intervention addresses a particular need, is likely to be acceptable, and the feasibility of delivery etc. The agility and speed with which RAP can be undertaken make it particularly useful when a problem has been identified to rapidly determine potential refinements to an intervention or programme and/or its delivery. Where differences in implementation have been identified then qualitative methods can explore reasons for 'positive deviance' to develop hypotheses about what has allowed the intervention or programme to succeed in some settings/participants when it has failed in the majority. Qualitative methods can be used to generate hypotheses about how a programme or intervention might work, particularly when, for example, a realist approach is taken and context-mechanism-outcome configurations developed (Manzano, 2016; Pawson and Tilley, 2004). This can usefully inform the design of future evaluation activities, including identifying relevant outcome measures.

Where there is a greater understanding of the type of intervention that is to be implemented methods such as A/B testing can be used to refine the intervention before wider scale up and testing. A/B testing is most suited to testing small modifications to a programme's design or messaging, where the changes introduced are intended to result in immediate change (Optipedia, n.d.). The focus on short-term outcomes, such as use and uptake, enables rapid testing of elements of a programme within a relatively short time frame but does not provide insight on longer-term impacts. As such A/B testing is particularly useful at the design or pilot stage of a programme and for answering questions about the early stages of a programme's theory of change. A/B testing has been used in South Africa to examine the impact of advertising content on demand for loans (Bertrand et al., 2010) and in Pakistan, Turkey, South Africa, Jordan, Bolivia, Peru and the Philippines to study the impact of varying message content of financial products in (Dibner-Dunlap and Rathore, 2016; Karlan et al., 2016). To be most effective A/B tests rely on good quality routine or administrative data and requires a large sample size to be able to measure small incremental changes.

Identify problems

We illustrate two example of quantitative methods for identifying problems; statistical process control (SPC), which combines time series analysis with graphical presentation of data, and bottle neck analysis, which identifies blocks in the implementation process. Qualitative methods are also important in highlighting unintended or unanticipated consequence of existing interventions.

SPC originates from manufacturing and has been used for monitoring and quality improvement in healthcare. It is a statistical method that combines time series analysis methods with graphical presentation of data to identify if observed variation in an outcome

deviates from the expected level of variations (Benneyan et al., 2003; Fereday, 2015). SPC is undertaken continually throughout a programme using data collected at standard intervals provided routine or operational data is available. It does not rely on reaching a pre-specified sample size as the statistical limits are varied accordingly; limits are adjusted when there is reason to believe that current limits are not appropriate to provide adequate signals for action. This means that SPC is able to detect process changes and trends from an early stage in the programme and that different outcome measures can be tracked overtime. The review did not identify examples of SPC having being used in a development context.

SPC is useful in situations where the context is complex and changeable as new outcomes can be dropped or added to the analysis as the intervention or programme is modified and its underpinning theory of change evolves. A highly adaptive approach to programming is likely to increase the number of outcome indicators that are measured. Changing outcomes is possible provided they are already available or easy to add to existing data collection tools. Where new data has to be collected this may have cost implications. SPC can also be used to detect potential differences arising from different implementation strategies between sites. This can highlight important differences that might warrant further investigation for example using qualitative methods to explore positive deviants.

Bottleneck analysis is one of three similar approaches to identifying the 'component(s) of a system that limits the overall performance or capacity' (O'Connell and Sharkey, 2013; Rio et al., 2015). Two related ideas are cascade analysis and community or systems effectiveness (Dellicour et al., 2016; Garnett et al., 2016; Webster et al., 2013). In each case a number of steps that link the population intended to benefit from an intervention and the population that do benefit are identified and assessed. Each step is conditional on the previous one having been met and only the population left at the end of all the steps would be anticipated to have achieved the desired outcome. The relative size of the population lost at

each step might indicate where the most urgent action is needed. For example, a bottleneck analysis of maternal and newborn health interventions in rural areas of the United Republic of Tanzania, found the largest bottleneck in one region was the availability of equipment, drugs and human resources in the facility, while in another the largest bottleneck was clinical practice (Baker et al., 2015). These methods are usefully combined with qualitative approaches to explore why the bottleneck has occurred and identify potential modifications to a programme.

Bottleneck analysis assumes a linear process; that achieving one step is a necessary condition to achieving the next. This implies that the hypothesised theory of change is the only route through which change can occur. To assess if this assumptions holds, requires an understanding of whether the population in one stage is the same as the population in the next, to ascertain whether it is a 'necessary' condition or whether other steps, not captured in the theory of change, might be sufficient to achieve the desired change (Davies, 2014). The analysis could be adapted to reflect changes in understanding of necessary and sufficient conditions and as the programme's theory of change evolves, provided data is available on the relevant outcomes.

Such analyses are often undertaken at a single point in time and provide a snap shot of need. Where routine or programme data is available the analysis can be undertaken relatively rapidly and could be repeated to assess whether the bottlenecks identified and size change overtime.

Test potential solutions

Experimental methods are used to assess the effectiveness of interventions or programmes and to ascertain causal relationships. Recent innovations including adaptive randomised control trials (RCTs) and modified stepped wedge trials present real opportunities for these

methods to usefully support timely approach to evaluation. Their use for complex interventions in international development however, has been highly restricted to date. The review identified one protocol for an adaptive RCT and one protocol for a modified stepped wedge trial in international development settings (Choko et al., 2017; Wechsberg et al., 2017).

Adaptive RCTs can be used to test multiple interventions in parallel before applying stopping rules as the evidence stacks up. This method may be particularly useful where it is not clear which interventions are most likely to be effective to achieve similar outcomes. The design includes multiple rounds of interim analysis that allows interventions that are not performing according to predetermined criteria to be terminated (Bothwell et al., 2018; Kairalla et al., 2012; Mahajan and Gupta, 2010). In addition to starting or stopping interventions modifications can include: adjusting the study population and sample size; and outcome-adaptive randomisation in which treatment allocation is skewed to those treatments that appear to be doing better. Potential modifications, and the criteria for implementing changes, need to be pre-specified based on decision rules in the study protocol.

The inclusion of a period of 'reflection' between each step of implementation in a modified stepped wedge trial makes this method useful where the basic form of an intervention has been decided upon at the outset but enables testing of the acceptability, feasibility and effectiveness of the intervention as it is implemented. Between steps formative research, including surveys, IDIs and FGDs, assess the acceptability and feasibility of implementing the intervention or programme and, where relevant, identify a revised plan to be implemented in the next step. At the end of the study it would be possible to compare the effect of the overall package of interventions on the pre-specified outcomes as in the original study, but additionally provides an evidence-based refined delivery plan for roll-out in other areas.

Both methods can be combined with methods such as SPC to determine whether causal mechanisms are being activated as anticipated as well as qualitative methods to understand the mechanism by which an intervention has impact, capture unanticipated outcomes and/or the influence of context (Stetler et al., 2006). The value of adapted or modified trials lies in their ability to make adjustments to the intervention or trial design as data is being collected, without undermining the validity or integrity of the study (Bhatt and Mehta, 2016; Bothwell et al., 2018; Kairalla et al., 2012; Korn and Freidlin, 2017; Lang, 2011; Thorlund et al., 2018; Villar et al., 2017). This provides both ongoing learning during the programme and confirmatory learning at the end of the trial, which could be generalised to other settings. Such designs require significant investment and expertise, can increase trial complexity and require sophisticated statistical techniques for the analysis.

Explain outcomes

Explaining outcomes draws primarily on qualitative methods to gather stakeholder and beneficiaries' perceptions of interventions and programmes or elucidation of their causal mechanisms. Examples include most significant change (MSC) and qualitative impact assessment protocol (QuIP). Both methods are undertaken retrospectively when sufficient time is anticipated to have passed to warrant examination of impact of an intervention or programme. The methods start by assessing whether meaningful change has occurred and work backwards to determine whether change can be attributed to the specific intervention (Beach and Pedersen, 2013; Lacouture et al., 2015).

MSC was originally developed as a form of participatory impact monitoring (Davies, 1996), to be used in a decentralised and participatory rural development programme, where standardised pre-defined indicators would not work. In each reporting period (initially 3 months), programme participants were asked to identify what they thought was the most

significant change, and its consequences. Stakeholder panels review these stories to identify the most significant and the consequences for the NGO's future work. In the decades since then MSC has been used in a wide variety of programmes, for both evaluation and monitoring purposes. Many different selection structures have been designed, to fit the different kinds of programmes and stakeholders involved (Davies and Dart, 2005). MSC is particularly valuable in highly complex settings where it is not known which activities are likely to have led to change and where causal mechanisms have either not been articulated at the project outset or cannot be agreed upon between stakeholders.

QuIP assesses impact through narrative causal statements from programme or intervention intended beneficiaries. The QuIP takes on the challenge of achieving sufficient credibility using timely qualitative methods in a way that can be both confirmatory (testing a theory of change) and exploratory (open to the unanticipated drivers and outcomes) (Copestake, 2014). It was developed through a grant to evaluate rural livelihood adaptation projects in Malawi and Ethiopia but has since been used to conduct relatively rapid studies in many other fields, including assessment of the social impact of ongoing programmes to promote decent work in Mexico, community self-organisation in Uganda and improved housing in India. (Copestake et al., 2018b; Copestake and Remnant, 2015). The QuIP incorporates features of a range of other qualitative approaches, including contribution analysis, process tracing, outcome harvesting and realist evaluation. It builds on ongoing quantitative monitoring of key indicators using semi-structured interviews and focus group discussions. It's potential as a timely and flexible approach is enhanced by requiring neither a baseline nor a comparison group. But like other forms of contribution analysis it tests the existence of causal pathways, but does not generate estimates of the magnitude of causal effects. Field data generated on drivers of change is open-ended and exploratory, because the field team is deliberately not informed of project theory (or even the identity of the project

being evaluated). But a critical part of the job of the analyst is to code the drivers of change identified according to whether they do explicitly or implicitly align with project theory or not. The QuIP aims to addresses the challenges of confirmation bias (where what people say is framed by how they are interviewed and possibly influenced by what they think you want to hear) through "blindfolding" interviewers and respondents from knowing the full details of the intervention evaluated (Copestake et al., 2018a).

These methods are generally undertaken at a single point in time, although they can be repeated to examine how perspectives change over time; in this way these method can assess both short and longer-term outcomes and can provide insights into whether a programme is having its intended impact and which activities are responsible for any observed change. These methods are particularly valuable where the interventions being implemented or the context are highly complex and changeable. They are also valuable where evaluation has not been incorporated from a programme's outset.

Both methods have the potential to be used for hypothesis testing, they examine what was achieved and how, to understand the relative importance of different activities undertaken. However, there is considerable flexibility as data collection is not restricted to pre-specified outcomes. This allows evaluators to capture unexpected outcomes and mechanisms of action, and can lead to new hypotheses and theories being generated. The timeliness of evidence can also be enhanced (relative to more traditional methods of qualitative research) by adopting more structured protocols for data coding, analysis and visualisation. The QuIP method has sought to speed up the process of synthesis and reporting by speeding up data analysis and reporting through use of bespoke spreadsheets, and interactive dashboards to supplement more formal reports.

Discussion

We set out to develop a framework to identify, categorise and critically appraise methods that can support a more timely approach to evaluation of international development programmes. We identified both quantitative and qualitative methods that can be used for different purposes, namely: supporting design, identifying problems and testing and explaining solutions. We suggest methods are selected based upon the purpose of the evaluation. This analysis highlights that different methods can fulfil multiple purpose; the particular method to be used should be selected based on the specific time-needs and flexibility of the programme.

Our review found there to be a dearth of examples of the application of methods being explicitly used for more timely approaches to evaluating international development programmes. Reasons for this may include that those conducting such evaluations rarely disseminate their findings through peer reviewed publications or through widely accessible grey literature. We are optimistic that there is significant potential for timely evaluation to improve international development outcomes. Realising this challenge however will require further understanding of a number of core issues and further work to develop and test methods to be used for timely evaluations. We reflect on some key issues that were repeatedly raised in discussions and in the literature.

To detect 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 run the risk of falsely detecting treatment effects or prematurely discarding promising interventions that do not show an impact at an early stage. It is therefore important to recognise 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).

The advantage of methods like adaptive and modified trials is that they can also provide confirmatory learning at the end of the trial, demonstrating whether an intervention had the intended impact by measuring pre-defined outcomes over the entire course of the trial. Outcomes are selected based on hypothesised causal chains. 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 overtime, 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).

We did not identify any documentation of the impact that measuring and basing decision on shorter term outcomes has in this setting through the literature review. However, during the symposium concerns were raised that these approaches might cause researchers to become too focused on short term outcomes at the expense of the longer term impacts and the impact on rigour. More research is needed to understand the validity and rigour of using more timely methods compared to endline analysis. This could be tested for example in a trial with different forms of timely evaluation as the different arms, for example, different timings of feeding back results, with different data sources informing the results.

Using pre-existing data can reduce the time and resources needed for quantitative methods. However, many development programmes have weak monitoring systems which make them less likely to be easily evaluable. Timeliness for many of the methods will therefore depend on the ability to collect, process and analyse data in a timely fashion. The challenge is to better leverage time series data from service delivery platforms and to make

such data useful (i.e. captures relevant outcome indicators in a timely manner) and of sufficient quality (i.e. measures needed to enhance completeness and accuracy of data).

The ability of routine data to respond to shifting priorities over time and the amount of time required for data collection and analysis, is variable depending upon the scale and ownership of the data collection system. While changes to the indicators in national-level routine systems are a major undertaking, other forms of routine data capture, such as programme monitoring data, may be more flexible and outcomes measured could be adjusted over time. The key therefore is in the initial design and whether an expectation of the need for flexibility has been built into the system. Where high quality routine data is available, then analysis is generally very rapid.

In settings where routine data is not available, innovative approaches to accessing routine data offer real potential (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). While, during the 2013-16 Ebola outbreak in West Africa real-time data surveys were undertaken resulting in significant lessons learned on the rapid collection, coordination and use of large amounts of data using new technologies and on coordination of this data amongst partners (Cori et al., 2017). The analysis of big data is already common place in the private sector; used for consumer profiling, personalised services and predictive analysis being used for advertising (UN Global Pulse, 2012). Technology that offers increasing opportunities for real time data analytics and their application should be explored more in development programmes.

The general consensus from the public symposium and literature review was that the use of mixed methods should be encouraged; quantitative approaches should be complemented for their interpretation, by process data, which is often qualitative. Mixing of methods can ensure a greater sensitivity amongst evaluators towards the potential threats to

the validity of conclusions (Ton, 2012). It has become a general expectation that impact evaluations be accompanied by a process evaluation and a similar approach makes perfect sense when considering timely evaluation within an ongoing programme.

A mixed methods approach may involve using complementary methods of data collection, but may also mean mixing or combining of theories, hypotheses, analyses and conceptual or analytical frameworks (Bamberger, 2012). Innovative approaches to mixing methods, stemming from the field of political science, have recently been proposed. Goertz's 'research triad' is a multi-methods approach which links not just quantitative (cross-case) with qualitative (within-case) inferences, but adds a third approach of the elucidation of causal mechanisms through for example, process tracing (Goertz and Mahoney, 2012). Amongst the qualitative approaches the interpretative approaches tend to have a focus on for example, the influence of power and the meaning behaviours, whilst a subset of methods are concerned with causal inference, mechanisms and generalisation (Goertz and Mahoney, 2012).

Stakeholder engagement is essential to ensure efficient incorporation of learning from timely evaluation into programme adaptations that can successfully be implemented. This can increase the utility of an evaluation to support programme improvement – an approach espoused by Patton called 'utilization focused evaluation' (Patton, 2008), in which end-users are identified and engaged from outset to guide other decisions that are made about the evaluation process. This has great benefits, though it also requires sufficient time and resources, as well as willingness on the part of the stakeholders. Evaluation also needs to be responsive such that results are available whilst there is momentum and engagement amongst staff. Sometimes staff may have solved problems that the evaluation later highlights the presence of, and therefore the evaluation is no longer relevant for pushing programme improvement.

The programmes within which the timely evaluation framework and approaches are applied

There is a close link between what the evaluation methods are trying to do, and the ability of programmes to incorporate and act on what they tell us either at programme outset, through adaptations over time that are responsive to monitoring data, or in acting on the results of comparative or explanatory studies on programme options or performance. A central issue to these are the intersection between programming flexibility / adaptability and the timing with which data from evaluation is "received" and how this links to programming cycles.

It was argued at the symposium that programme improvement is only really possible when: 1) programmes are small; 2) there is a specific intention to learn and adapt; 3) when results are immediately available; 4) when changes to the programme are small-scale within the capacity of the programme to deliver; and 5) when programmes have time to try out various options before rolling out to reach a large number of beneficiaries (Aly Visram personal communication). Large scale improvements are difficult if not impossible to implement, especially because they require significant investment. Large scale improvements are also likely to be beyond the financial capacity of programmes that have pre-budgeted based on a fixed plan of action. The proposition of achievement through small incremental changes is supported by the idea from evolutionary theory of 'the adjacent possible' (Srivastava, 2014).

Effective use of data requires appropriate data, that reaches the right people, who understand the data as presented, are able to transform it as required, and have the power to make decisions or have access to those who do. The guidance on change must then be produced and transferred back to implementers who are able, and willing, to put changes into action. The presence of programme and institutional structures required to support this process, which in itself is complex, will vary.

Uncertainty over what evidence might be needed and when, is often compounded by delays in the time it takes commissioners and evaluators to respond. Empirical evidence on the processes involved in generating evidence is lacking, partly perhaps because the scope for generalising usefully about it is limited by context-specificity. Having set out to develop a more agile approach to collecting 'good enough' evidence in the form of the 'QuIP' James Copestake reflected at the public Symposium, on practical obstacles to doing so.

Starting with the demand side, delays arise in securing agreement on the design, budget, release of sample-frame data, clarity on the theory of change needed to guide data coding and on obtaining ethics approval sometimes across more than one institution. These are particularly likely when the commissioner seeking an evaluation and the organisation executing the activity being evaluated are distrustful of each other. Delays arise from variation in the nature of the primary intended audience and their expectation of what evidence should look like, which may range from a flexible data dashboard to a glossy report. The more controversial the findings (and hence perhaps the more important) the more the likelihood of lengthy negotiation over an 'acceptable' final draft. Meanwhile, on the supply side, the challenge of mobilising appropriate and available staff for data collection is often compounded by problems securing permission to enter the field, finalising contracts and securing ethical approval (Gamble, 2006; Patton, 2013; Portela et al., 2015).

There is a need to test the scope of timely evaluation methods and to determine which programmes they can or should be applied to. There is limited evidence in particular for outcome evaluation methods presented here (adapted RCTs and modified stepped wedge trials), which might support large scale testing and change.

Assessing the impact of timely evaluation

Timely evaluation approaches are likely to be more time and resource intensive. All of the

methods presented are likely to be resource intensive and require more data to be collected than traditional evaluation methods. Methods that do not test a specific causal mechanism need to capture a wider range of outcomes and casual pathways. Whilst, methods that aim to rapidly test changes or compare multiple-interventions rely on ongoing or repeat measurement of data. The methods are anticipated to represent overall value for money as they result in the programme having a higher chance of success. However the impact/benefits of undertaking more timely approaches to evaluation are not well understood (O'Donnell, 2016). There is therefore a need to determine whether undertaking a timely evaluation does lead to greater impact that traditional approaches and represent value for money.

It is important to understand the implications of learning more for this time on our ability to learn more for next time. Where an intervention changes over time there is a need to identify when it becomes an entirely new intervention and to recognise when the use of these methods become an intervention in themselves (Portela et al., 2015). If this is the case the use of these methods may need to be incorporated into interventions being replicated in different settings. It is questionable then whether we can learn anything on scaling up or replication in other settings using these approaches. It is necessary to understand the nature of implementation and the degree to which evaluation activities influence and contribute to the overall results of a programme.

Limitations of our approach

There were several limitations to our approach. Our scoping seminar and public symposium were interesting and exciting events, which provided an opportunity for broad discussion of timely evaluation within international development. Although in setting the agenda and selecting speakers we attempted to focus some of the discussions, the topic was new for many participants and therefore the discussions quite broad.

Reviewing the literature on this topic proved to be extraordinary difficult due to the wide range of terminologies around timely evaluation, programme improvement and adaptive learning. Many of the methods we identified were specific to certain niches for example, quality improvement initiatives. There were also a range of terminologies for what in effect were very similar methods. In addition to problems in terminology, there were many examples of methods being advocated for and described without any examples of their practical application or critique of this application.

Although we attempted to embrace a wide range of sectors in our paper, the experience of the majority of the author team, and participants of the scoping session and public symposium is in the health sector and therefore most of our examples are from the health sector. We hope however, that our framework and discussion of approaches and methods will provide a starting point, which can be applied across sectors.

Identification, categorisation and better selection of methods for timely evaluation within specific programmes can only go so far in improving outcomes: uncertainty will always remain about "what works, for whom and under what circumstances". Borrowing this mantra from the tradition of realist evaluation is not an accident because a complexity ontology is what underpins it, and its recognition that evaluation is unavoidably political as well as technical (Pawson, 2013).

Recommendations for further research

Based on our discussions and review of the literature we recommend further research on timely evaluation including:

Testing and development of framework. The framework should be tested to ensure fit for purpose. Workshops convening relevant stakeholders including researchers, implementers and decision makers could assess the utility of the framework for selecting

methods and determining the optimum mix of methods for addressing different development projects being conducted in different contexts and settings. Through testing would also identify research priorities for developing new or adapting existing methods to meet the needs of a more timely approach to evaluation.

Developing guidelines and best practices. The framework should be developed further to provide guidance on best practices on timely evaluation for programme improvement for different types of projects within different contexts. This would involve formulating a matrix of recommended methods with guidance on their applicability for different projects, contexts and sectors, for example, education and agriculture.

Evaluating adaptive management interventions. While the flexible approaches underlying adaptive management are very promising, these remain to be rigorously evaluated.

Conducting adaptive trials. The application of adaptive trials to multi-component interventions where different packages of configurations are tested, where there are ethical issues and decisions have to be made quickly. For example, humanitarian assistance interventions would be one of such cases.

Conclusion

There is significant potential for more timely evaluation to improve international development outcomes. Despite the availability of new approaches to mixing and adapting existing methods and the potential for new technologies to enhance data collection, there is a dearth of examples of their application. Research is needed to provide an empirical evidence base upon which to further develop and appraise the application of these methods, across sectors and contexts within international development.

References:

Optipedia, n.d. A/B Testing [WWW Document]. Optimizely. URL https://www.optimizely.com/optimization-glossary/ab-testing/ (accessed 1.2.18).

Andrews, M., 2015. Explaining positive deviance in public sector reforms in development. World Development 74, 197–208.

Baker, U., Peterson, S., Marchant, T., Mbaruku, G., Temu, S., Manzi, F., Hanson, C., 2015. Identifying implementation bottlenecks for maternal and newborn health interventions in rural districts of the United Republic of Tanzania. Bull. World Health Organ. 93, 380–389. https://doi.org/10.2471/BLT.14.141879

Bamberger, M., 2012. Introduction to mixed methods in impact evaluation. Impact Evaluation Notes 3, 1–38.

Barder, O., Ramalingam, B., 2012. Complexity, Adaptation, and Results. Center For Global Development.

Barnett, C., Munslow, T., 2014. Process Tracing: The Potential and Pitfalls for Impact Evaluation in International Development. Summary of a Workshop held on 7 May 2014 (No. IDS Evidence Report 102). Institute of Development Studies.

Barr, J., 2015. Monitoring and Evaluating Flexible and Adaptive Programming. Itad.

Beach, D., Pedersen, R.B., 2013. Process-Tracing Methods: Foundations and Guidelines. University of Michigan Press.

Beebe, J., 2001. Rapid Assessment Process: An Introduction. Rowman Altamira.

Befani, B., 2013. Between complexity and generalization: Addressing evaluation challenges with QCA. Evaluation 19, 269–283. https://doi.org/10.1177/1474022213493839

Befani, B., Mayne, J., 2014. Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation. IDS Bulletin 45, 17–36. https://doi.org/10.1111/1759-5436.12110

Benneyan, J.C., Lloyd, R.C., Plsek, P.E., 2003. Statistical process control as a tool for research and healthcare improvement. BMJ Quality & Safety 12, 458–464. https://doi.org/10.1136/qhc.12.6.458

Lopez Bernal, J., Cummins, S., Gasparrini, A., 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 46, 348–355. https://doi.org/10.1093/ije/dyw098

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., Zinman, J., 2010. What's advertising content worth? Evidence from a consumer credit marketing field experiment. The Quarterly Journal of Economics 125, 263–306.

Bhatt, D.L., Mehta, C., 2016. Adaptive Designs for Clinical Trials. N. Engl. J. Med. 375, 65–74. https://doi.org/10.1056/NEJMra1510061

Biglan, A., Ary, D., Wagenaar, A.C., 2000. The value of interrupted time-series experiments for community intervention research. Prev Sci 1, 31–49.

Bothwell, L.E., Avorn, J., Khan, N.F., Kesselheim, A.S., 2018. Adaptive design clinical trials: a review of the literature and ClinicalTrials.gov. BMJ Open 8, e018320. https://doi.org/10.1136/bmjopen-2017-018320

Burke, L.E., Shiffman, S., Music, E., Styn, M.A., Kriska, A., Smailagic, A., Siewiorek, D., Ewing, L.J., Chasens, E., French, B., Mancino, J., Mendez, D., Strollo, P., Rathbun, S.L., 2017. Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges. J Med Internet Res 19. https://doi.org/10.2196/jmir.7138

Busza, J., Teferra, S., Omer, S., Zimmerman, C., 2017. Learning from returnee Ethiopian migrant domestic workers: a qualitative assessment to reduce the risk of human trafficking. Global Health 13. https://doi.org/10.1186/s12992-017-0293-x

Butler, L.M., 1995. The Sondeo: a rapid reconnaissance approach for situational assessment.

Cellamare, M., Ventz, S., Baudin, E., Mitnick, C.D., Trippa, L., 2017. A Bayesian response-adaptive trial in tuberculosis: The endTB trial. Clin Trials 14, 17–28. https://doi.org/10.1177/1740774516665090

Choko, A.T., Fielding, K., Stallard, N., Maheswaran, H., Lepine, A., Desmond, N., Kumwenda, M.K., Corbett, E.L., 2017. Investigating interventions to increase uptake of HIV testing and linkage into care or prevention for male partners of pregnant women in antenatal clinics in Blantyre, Malawi: study protocol for a cluster randomised trial. Trials 18. https://doi.org/10.1186/s13063-017-2093-2

Connors, S.C., Nyaude, S., Challender, A., Aagaard, E., Velez, C., Hakim, J., 2017. Evaluating the Impact of the Medical Education Partnership Initiative at the University of Zimbabwe College of Health Sciences Using the Most Significant Change Technique. Acad Med 92, 1264–1268. https://doi.org/10.1097/ACM.0000000000001519

Copestake, J., 2014. Credible impact evaluation in complex contexts: Confirmatory and exploratory approaches. Evaluation 20, 412–427. https://doi.org/10.1177/1356389014550559

Copestake, J., Allan, C., Bekkum, W. van, Belay, M., Goshu, T., Mvula, P., Remnant, F., Thomas, E., Zerahun, Z., 2018a. Managing relationships in qualitative impact evaluation of international development: QuIP choreography as a case study. Evaluation 24, 169–184. https://doi.org/10.1177/1356389018763243

Copestake, J., Morsink, M., Remnant, F. (Eds.), 2018b. Attributing development impact: the QuIP case book. Practical Action, Rugby.

Copestake, J., Remnant, F., 2015. Assessing Rural Transformations: Piloting a Qualitative Impact Protocol in Malawi and Ethiopia, in: Mixed Methods Research in Poverty and Vulnerability. Palgrave Macmillan, London, pp. 119–148. https://doi.org/10.1057/9781137452511_6

Cori, A., Donnelly, C.A., Dorigatti, I., Ferguson, N.M., Fraser, C., Garske, T., Jombart, T., Nedjati-Gilani, G., Nouvellet, P., Riley, S., Van Kerkhove, M.D., Mills, H.L., Blake, I.M., 2017. Key data for outbreak evaluation: building on the Ebola experience. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 372. https://doi.org/10.1098/rstb.2016.0371

Davies, R., 2016a. Qualitative Comparative Analysis [WWW Document]. Better Evaluation. URL http://www.betterevaluation.org/en/evaluation-options/qualitative_comparative_analysis (accessed 1.3.18).

Davies, R., 2016b. Evaluating the impact of flexible development interventions using a 'loose' theory of change Reflections on the Australia-Mekong NGO Engagement Platform (A method lab publication). Overseas Development Institute, London.

Davies, R., 2014. Thinking about set relationships within monitoring data. Rick on the Road.

Davies, R., 1996. An evolutionary approach to facilitating organisational learning: An experiment by the Christian Commission for Development in Bangladesh. University of Swansea, Centre for Development Studies.

Davies, R., Dart, J., 2005. The "Most Significant Change" (MSC) Technique: A Guide to Its Use.

Davies, R., Laidlaw, J., Rogers, P., 2016. Process Tracing [WWW Document]. Better Evaluation.

DDD, 2014. Doing Development Differently [WWW Document]. Doing Development Differently. URL http://doingdevelopmentdifferently.com/the-ddd-manifesto/ (accessed 1.17.18).

Dellicour, S., Hill, J., Bruce, J., Ouma, P., Marwanga, D., Otieno, P., Desai, M., Hamel, M.J., Kariuki, S., Webster, J., 2016. Effectiveness of the delivery of interventions to prevent malaria in pregnancy in Kenya. Malar. J. 15, 221. https://doi.org/10.1186/s12936-016-1261-2

DFID, 2012. Results in Fragile and Conflict-Affected Stated and Situations: How to Note. Department for International Development.

Dibner-Dunlap, A., Rathore, Y., 2016. Beyond RCTs: How Rapid-Fire Testing Can Build Better Financial Products [WWW Document]. Innovations for Poverty Action. URL https://www.poverty-action.org/blog/beyond-rcts-how-rapid-fire-testing-can-build-better-financial-products (accessed 1.2.18).

Earl, S., Carden, F., Smutylo, T., 2001. Outcome Mapping: Building Learning and Reflection into Development Programs. International Development Research Centre, Ottawa, Canada.

Eirich, F., Morrison, A., n.d. Guide 6: Contribution Analysis, Social Sciene Methods Series. Scottish Government.

Fereday, S., 2015. A guide to quality improvement methods. Healthcare Quality Improvement Partnership.

Gamble, J., 2006. A Developmental Evaluation Primer. The J.W. McConnell Family Foundation, Canada.

Ganann, R., Ciliska, D., Thomas, H., 2010. Expediting systematic reviews: methods and implications of rapid reviews. Implementation Science 5, 56. https://doi.org/10.1186/1748-5908-5-56

Garnett, G.P., Hallett, T.B., Takaruza, A., Hargreaves, J., Rhead, R., Warren, M., Nyamukapa, C., Gregson, S., 2016. Providing a conceptual framework for HIV prevention cascades and assessing feasibility of empirical measurement with data from east Zimbabwe: a case study. Lancet HIV 3, e297-306. https://doi.org/10.1016/S2352-3018(16)30039-X

Goertz, G., Mahoney, J., 2012. A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences. Princeton University Press.

Grant, M.J., Booth, A., 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal 26, 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Green, D., 2015. Doing Development Differently: a great discussion on Adaptive Management (no, really). From Poverty to Power.

Harris, K.J., Jerome, N.W., Fawcett, S.B., 1997. Rapid assessment procedures: A review and critique. Human Organization 56, 375–378.

HEARD Project, 2018. Rapid Review vs. Systematic Review: What are the differences? [WWW Document]. USAID. URL https://www.heardproject.org/news/rapid-review-vs-systematic-review-what-are-the-differences/ (accessed 10.26.18).

Hildebrand, P.E., 1981. Combining disciplines in rapid appraisal: the Sondeo approach. Agricultural Administration 8, 423–432.

Ho, L.S., Labrecque, G., Batonon, I., Salsi, V., Ratnayake, R., 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. Confl Health 9, 27. https://doi.org/10.1186/s13031-015-0055-4

Hubbard, B., 2010. Root Cause Analysis (Overview). Lean Learning Revolution!

IPA, 2016. Introduction to Rapid-Fire Operational Testing for Social Programs (Goldilocks Deep Dive). Innovations for Poverty Action.
Jones, H., Hearn, S., 2009. Outcome Mapping: A realistic alternative for planning, monitoring and evaluation (Working and discussion paper). Overseas Development Institute.
Jordan, E., Gross, M.E., Javernick-Will, A.N., Garvin, M.J., 2011. Use and misuse of qualitative comparative analysis. Construction Management and Economics 29, 1159–1173. https://doi.org/10.1080/01446193.2011.640339
Kairalla, J.A., Coffey, C.S., Thomann, M.A., Muller, K.E., 2012. Adaptive trial designs: a review of barriers and opportunities. Trials 13, 145. https://doi.org/10.1186/1745-6215-13-145
Kane, H., Lewis, M.A., Williams, P.A., Kahwati, L.C., 2014. Using qualitative comparative analysis to understand and quantify translation and implementation. Transl Behav Med 4, 201–208. https://doi.org/10.1007/s13142-014-0251-6
Karlan, D., 2017. Nimble RCTs. A Powerful Methodology in the Program Design Toolbox. Innovations for Poverty Action, Yale University.
Karlan, D., McConnell, M., Mullainathan, S., Zinman, J., 2016. Getting to the top of mind: How reminders increase saving. Management Science 62, 3393–3411.
Kontopantelis, E., Doran, T., Springate, D.A., Buchan, I., Reeves, D., 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ 350, h2750.
Korn, E.L., Freidlin, B., 2017. Adaptive Clinical Trials: Advantages and Disadvantages of Various Adaptive Design Elements. J Natl Cancer Inst 109. https://doi.org/10.1093/jnci/djx013
Lacouture, A., Breton, E., Guichard, A., Ridde, V., 2015. The concept of mechanism from a realist approach: a scoping review to facilitate its operationalization in public health program evaluation. Implement Sci 10, 153. https://doi.org/10.1186/s13012-015-0345-7
Ladner, D., 2015. Strategy Testing: An Innovative Approach to Monitoring Highly Flexible Aid Programs (Case Study no 3), Working Politically in Practice. The Asia Foundation.
Lang, T., 2011. Adaptive trial design: could we use this approach to improve clinical trials in the field of global health? Am. J. Trop. Med. Hyg. 85, 967–970. https://doi.org/10.4269/ajtmh.2011.11-0151
Limato, R., Ahmed, R., Magdalena, A., Nasir, S., Kotvojs, F., 2018. Use of most significant change (MSC) technique to evaluate health promotion training of maternal community health workers in Cianjur district, Indonesia. Eval Program Plann 66, 102–110. https://doi.org/10.1016/j.evalprogplan.2017.10.011
Lopez Bernal, J., Cummins, S., Gasparrini, A., 2018. The use of controls in interrupted time series studies of public health interventions. Int J Epidemiol. https://doi.org/10.1093/ije/dyy135
Mahajan, R., Gupta, K., 2010. Adaptive design clinical trials: Methodology, challenges and prospect. Indian J Pharmacol 42, 201–207. https://doi.org/10.4103/0253-7613.68417
Manderson, L., Aaby, P., 1992. Can rapid anthropological procedures be applied to tropical diseases? Health policy and planning 7, 46–55.
Manzano, A., 2016. The craft of interviewing in realist evaluation. Evaluation 22, 342–360. https://doi.org/10.1177/1356389016638615
Mayne, J., 2008. Contribution Analysis [WWW Document]. Better Evaluation. URL https://www.betterevaluation.org/en/plan/approach/contribution_analysis (accessed 10.26.18).

Moore, G.F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D., 2015. Process evaluation of complex interventions: Medical Research Council guidance. bmj 350, h1258.

O'Connell, T., Sharkey, A., 2013. Reaching universal health coverage: using a modified Tanahashi model sub-nationally to attain equitable and effective coverage. UNICEF, New York.

ODI, 2009. Strategy Development: Outcome Mapping, in: Tools for Knowledge and Learning: A Guide for Development and Humanitarian Organisations.

O'Donnell, M., 2016. Adaptive management: what it means for civil society organisations. Bond, London.

Patton, M.Q., 2013. Utilization-Focused Evaluation (U-FE) Checklist [WWW Document]. URL https://wmich.edu/sites/default/files/attachments/u350/2014/UFE_checklist_2013.pdf (accessed 12.12.17).

Patton, M.Q., 2008. Utilization-focused evaluation. Sage publications.

Pawson, R., 2013. The science of evaluation: a realist manifesto. Sage.

Pawson, R., Tilley, N., 2004. Realist Evaluation.

Peerally, M.F., Carr, S., Waring, J., Dixon-Woods, M., 2017. The problem with root cause analysis. BMJ Qual Saf 26, 417–422. https://doi.org/10.1136/bmjqs-2016-005511

Peters, A., 2018. At these camps, refugees can give real-time customer feedback [WWW Document]. Fast Company. URL https://www.fastcompany.com/40575160/at-these-camps-refugees-can-give-real-time-customer-feedback (accessed 7.24.18).

Portela, M.C., Pronovost, P.J., Woodcock, T., Carter, P., Dixon-Woods, M., 2015. How to study improvement interventions: a brief overview of possible study types. BMJ Qual Saf 24, 325–336. https://doi.org/10.1136/bmjqs-2014-003620

Positive Deviance Initiative, 2017. What is Positive Deviance? [WWW Document]. Positive Deviance Initiative. URL https://positivedeviance.org/ (accessed 3.20.18).

Research to Action, 2012. Outcome Mapping: A Basic Introduction [WWW Document]. Research to Action. URL http://www.researchtoaction.org/2012/01/outcome-mapping-a-basic-introduction/ (accessed 12.7.17).

Rio, D., Hedges, J., Woodhead, S., Rogers, E., 2015. What is the bottleneck analysis approach for the management of severe acute malnutrition? UNICEF and Action Against Hunger.

Schünemann, H. (Ed.), 2015. Advances in Rapid Reviews. Systematic Reviews 4.

Shiffman, S., Stone, A.A., Hufford, M.R., 2008. Ecological momentary assessment. Annu Rev Clin Psychol 4, 1–32.

Smutylo, T., 2005. Outcome Mapping: A method for tracking behavioural changes in development programs (No. ILAC Brief 7).

Srivastava, K., 2014. The 'Adjacent Possible' of Big Data: What Evolution Teaches About Insights Generation [WWW Document]. WIRED. URL https://www.wired.com/insights/2014/12/the-adjacent-possible-of-big-data/ (accessed 1.21.18).

Stetler, C.B., Legro, M.W., Wallace, C.M., Bowman, C., Guihan, M., Hagedorn, H., Kimmel, B., Sharp, N.D., Smith, J.L., 2006. The Role of Formative Evaluation in Implementation Research and the QUERI Experience. J Gen Intern Med 21, S1–S8. https://doi.org/10.1111/j.1525-1497.2006.00355.x

Talcott, F., Scholz, V., 2015. Methodology Guide to Process Tracing for Christian Aid. the International Non-Governmental Training and Research Centre, Oxford.

Tanahashi, T., 1978. Health service coverage and its evaluation. Bull World Health Organ 56, 295–303.

Theiss-Nyland, K., Koné, D., Karema, C., Ejersa, W., Webster, J., Lines, J., 2017. The relative roles of ANC and EPI in the continuous distribution of LLINs: a qualitative study in four countries. Health Policy Plan 32, 467–475. https://doi.org/10.1093/heapol/czw158

Thorlund, K., Haggstrom, J., Park, J.J., Mills, E.J., 2018. Key design considerations for adaptive clinical trials: a primer for clinicians. BMJ 360, k698. https://doi.org/10.1136/bmj.k698

Ton, G., 2012. The mixing of methods: A three-step process for improving rigour in impact evaluations. Evaluation 18, 5–25.

Tricco, A.C., Antony, J., Zarin, W., Strifler, L., Ghassemi, M., Ivory, J., Perrier, L., Hutton, B., Moher, D., Straus, S.E., 2015. A scoping review of rapid review methods. BMC Medicine 13, 224. https://doi.org/10.1186/s12916-015-0465-6

Tricco, A.C., Langlois, E., Straus, S.E., 2017. Rapid reviews to strengthen health policy and systems: a practical guide. World Health Organization, Alliance for Health Policy and Systems Research, Geneva.

UN Global Pulse, 2012. Big Data for Development: Challenges and Opportunities. United Nations.

Valters, C., Cummings, C., Nixon, H., 2016. Putting learning at the centre. Adaptive development programming in practice. Overseas Development Institute.

Villar, S.S., Bowden, J., Wason, J., 2017. Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends? Pharm Stat. https://doi.org/10.1002/pst.1845

Vlassoff, C., Tanner, M., 1992. The relevance of rapid assessment to health research and interventions. Health policy and planning 7, 1–9.

Walji, A., Vein, C., 2013. Learning from Data-Driven Delivery [WWW Document]. The World Bank. URL http://blogs.worldbank.org/voices/learning-data-driven-delivery (accessed 10.11.17).

Webster, J., Kayentao, K., Bruce, J., Diawara, S.I., Abathina, A., Haiballa, A.A., Doumbo, O.K., Hill, J., 2013. Prevention of malaria in pregnancy with intermittent preventive treatment and insecticide treated nets in Mali: a quantitative health systems effectiveness analysis. PLoS ONE 8, e67520. https://doi.org/10.1371/journal.pone.0067520

Wechsberg, W.M., Ndirangu, J.W., Speizer, I.S., Zule, W.A., Gumula, W., Peasant, C., Browne, F.A., Dunlap, L., 2017. An implementation science protocol of the Women's Health CoOp in healthcare settings in Cape Town, South Africa: A stepped-wedge design. BMC Womens Health 17. https://doi.org/10.1186/s12905-017-0433-8

White, H., 2013. Using the causal chain to make sense of the numbers [WWW Document]. International Initiative for Impact Evaluation. URL http://www.3ieimpact.org/en/announcements/2013/02/12/using-causal-chain-make-sense-numbers/ (accessed 10.17.18).

White, H., Phillips, D., 2012. Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework (Working Paper 15). International Initiative for Impact Evaluation.

Wilson-Grau, R., 2015. Outcome Harvesting [WWW Document]. Better Evaluation. URL http://www.betterevaluation.org/en/plan/approach/outcome_harvesting (accessed 1.4.18).

Wilson-Grau, R., Britt, H., 2012. Outome Harvesting Brief. Ford Foundation.

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, EASE '14.

ACM, New York, NY, USA, pp. 38:1–38:10. https://doi.org/10.1145/2601248.2601268

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, 1–14.

Table 1 Search terms

Search	Terms (title/abstract/key word)
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”
2	Humanitarian OR International Development
3	1 AND 2

Table 2 Approaches for timely evaluation and adaptive learning

Accountable aid, action research, active research, adaptive development, adaptive learning, adaptive management, adaptive programming, adaptive strategy, agile working practices, appreciative inquiry, augmented feedback, behaviour centred design / human centred design, better programme delivery, , citizen engagement, collaborating learning and adapting, complexity thinking, constituent voice, continuous evaluation, continuous improvement, creative design process, developmental evaluation, dynamic adaptive pathways, experiential learning, extrinsic feedback, feedback loops, feedback mechanisms, formative evaluation, iterative inquiry framework, iterative evaluation process, knowledge of results feedback, lean startup learning culture/system, model for improvement, nimble evaluations, performance management, plan-do-study-act cycle, problem driven iterative adaptation, problem based iterative adaptation, quality improvement, rapid assessment / rapid assessment process / rapid assessment methodology, rapid-cycle assessment, rapid cycle evaluation, rapid cycle quality improvement, rapid evaluation (and assessment) methods, rapid feedback evaluation, rapid qualitative enquiry, real time adaption, real time evaluation, social learning, strategy testing, utilisation focused evaluation

Table 3 Evaluation methods reviewed

Method	Description	Use and timing	Strengths/Weaknesses/ Considerations
A/B tests (also known as Nimble RCT, split tests, rapid-fire tests, bucket testing, randomized field experiments) (Dibner-Dunlap and Rathore, 2016; IPA, 2016; Karlan, 2017; Optipedia, n.d.)	Clinical study design; participants are randomly assigned to receive a variation of the same intervention. Compares the effect of the adaptations on short-term outcomes.	• Simultaneous testing of low-cost modifications to a programme's design or message, where changes are anticipated to result in immediate change. • Particularly useful at the design or pilot stage of a programme and for answering questions about the early stages of a programme's theory of change. • Focus on short-term outcomes and use of pre-existing data enables rapid testing of elements of a programme within a relatively short time frame.	• Focus on shorter-term outcomes such as uptake and use but does not provide insight on whether the changes had an impact on longer-term changes. • Small effect sizes as examining incremental change; requires large samples. • Relies on good quality routine/administrative data being available.
Adaptive randomised control trial (Bhatt and Mehta, 2016; Kairalla et al., 2012; Korn and Freidlin, 2017; Lang, 2011; Villar et al., 2017; Cellamare et al., 2017; Choko et al., 2017; Bothwell et al., 2018; Mahajan and Gupta, 2010; Thorlund et al., 2018)	Clinical study design; compares outcomes between control and intervention group. Outcomes are analysed at predefined interim time points and modifications to the study can be implemented based on the findings of the interim analysis. Modifications are made based on pre-specified decision rules.	• Where not clear which interventions are most likely to be effective to achieve similar outcomes, as allow simultaneous testing of multiple experimental arms. • Ongoing learning based on interim analysis: stop or start treatment arms; adjust the study population and sample size; skew treatment allocation to those treatments that appear to be doing better. • Provides confirmatory learning at end of trial. • Reduces time by combining trial phases into a single study.	• Ability to make adjustments to the intervention or trial design as data is being collected, without undermining the validity or integrity of the study. • Outcomes to be measured specified at trial outset. • Decisions made during trial based on interim-findings. • More resource intensive; requires interim data collection and more rounds of analysis than a classic RCT. • Increased trial complexity; requires sophisticated statistical techniques for the analysis. • Introducing new trial arms reduces statistical efficiency. • Potential for bias from temporal trends e.g. if participants recruited at early stages differ to those recruited at latter stages.

Method	Description	Use and timing	Strengths/Weaknesses/ Considerations
Bottleneck analysis/Cascade analysis/Community or systems effectiveness/Funnel of attrition (Davies, 2014; Dellicour et al., 2016; Garnett et al., 2016; O'Connell and Sharkey, 2013; Rio et al., 2015; Tanahashi, 1978; Webster et al., 2013; White, 2013)	Quantitative analysis. Identifies the steps that link the intended beneficiaries from the actual beneficiaries. Each step is conditional on the previous one having been met and only the population left at the end of all the steps have achieved the desired outcome. The relative size of the population lost at each step might indicate where the most urgent action is needed. Analysis can be stratified to understand differences between sub-groups.	• Identifies component(s) of a system that limits its overall performance or capacity. • Undertaken once an intervention is running and anticipate that an impact should have occurred. • Often undertaken at a single point in time providing a snap shot of need; where routine or programme data is available analysis could be repeated to assess whether the bottlenecks identified, and size, change overtime.	• Requires population level data; routine or programmatic survey data • Requires a hypothesised casual pathway; assumes achieving one step is a necessary condition to achieving the next e.g. the Theory of Change is the only route through which change can occur. Requires an understanding of whether the population in one stage is the same as the population in the next, to ascertain whether it is a 'necessary' condition or whether other steps, not captured in the Theory of Change, might be sufficient to achieve the desired change. • Casual pathways can be modified overtime. • Does not assess causality.
Contribution analysis (Befani and Mayne, 2014; Eirich and Morrison, n.d.; Mayne, 2008)	A structured approach to explore and estimate the relative contribution of an intervention to an outcome. Maps out ongoing activities that are being undertaken that are expected to contribute to a particular outcome. Collects diverse evidence to populate 'performance stories' against a pre-specified theory of change.	• Used to confirm or revise a theory of change. • Provide feedback on what is driving change and relative contribution of a particular intervention. • Particularly useful in situations where an experimental method is not feasible. • Best suited to large scale programmes	• Retrospective approach, little or no scope for varying how the programme is implemented. • Considers the relative impact of other activities on a desired outcome.
Ecological momentary assessment. Ambulatory Assessment/Experience Sampling (Burke et al., 2017; Shiffman et al., 2008)	Longitudinal design; method for collecting data in real-time, in real world settings. Participants complete short assessments on their current experiences / behaviours / moods / environment at multiple random moments over time. Two approaches: 1) signal-contingent recording – assessed a fixed number of times per day/week etc. on a random schedule; 2) event-contingent recording – assessed following exposure to specific events.	• Used to study psychological, behavioural, and physiological processes in the natural environment. • When using mobile technology data generated in real-time.	• Minimises recall bias; combines actual exposure measurements with momentary-measured outcomes. • Repeat sampling of same individuals allows for within- and between-participant analysis. Can examine causality between exposures and behaviours. • Challenges; logistic, analytic, and interpretation problems • Increasing availability of mobile technology offers increased utility

Method	Description	Use and timing	Strengths/Weaknesses/ Considerations
Interrupted time series analysis (Biglan et al., 2000; Kontopantelis et al., 2015; Lopez Bernal et al., 2018, 2017)	A quasi-experimental method (others include difference-in-difference, synthetic controls, matching, regression discontinuity); model trend in outcome before and after intervention is introduced against what would have happened is the intervention was not introduced. Any change in the level of the outcome or in the rate of change over time, compared to the model, can be interpreted as the effect of the intervention.	• To determine the effect of an intervention implemented at a specific time point in the absence of a parallel control. • More complex designs can be used in situations where intervention is stopped/reversed or with multi-component interventions where different steps are implemented at different time points.	• Requires a large amount of data to be collected before and after intervention is introduced at equally-spaced time intervals. • Outcomes to be measured need to be pre-specified. • Population under study act as own control; although analysis can also include a control group e.g. from a different area. • Requires programmatic or routine data.
Modified stepped wedge trials (Wechsberg et al., 2017)	Clinical study design; compares outcomes between control and intervention arms within each step. A modified design incorporates a period of reflection at the end of each step for example undertaking surveys/IDI/FGDs to understand how the intervention is working etc. Modifications to intervention can be implemented before the next step.	• Prospective; to test and adapt implementation strategies. • Ongoing learning; make sequential changes to the intervention. • Confirmatory learning at the end of trial possible to compare the effect of the overall package of interventions on the pre-specified outcomes as in the original study.	• Potential bias from temporal trends e.g. if participants recruited early in the trial differ to those recruited later. • Adaptations made during trial based on interim analysis. • Outcomes to be measured specified at trial outset; although additional unanticipated outcomes can be explored in the 'period of reflection'. • More resource intensive; requires additional data collection between steps; time needed to undertake data collection and analysis can increase length of trial. • Can increase trial complexity • Limited evidence of use from literature.
Most significant change (Connors et al., 2017; Davies and Dart, 2005; Ho et al., 2015; Limato et al., 2018; White and Phillips, 2012)	Participatory qualitative method; use qualitative methods to collect programme beneficiaries' stories of recent significant change in their lives and the key activities they think led to these changes. Panel of stakeholders select what they consider to be the most significant stories, to arrive at a reduced set of changes.	• Retrospective; undertaken when anticipate some impact should have occurred. • Can be undertaken on an ongoing basis throughout the project cycle to reveal changes in stakeholder's perspectives at different time points. • Useful in contexts where programme already running or highly complex setting and not clear what impact may have.	• Does not get at causality • Measures intermediate outcomes and programme impact. • Can capture unexpected outcomes as do not have to hypothesis causal pathways between activities and outcomes. • Stories collected at a single point in time so does not account for changes due to temporality.

Method	Description	Use and timing	Strengths/Weaknesses/ Considerations
			• Human resource intensive; collect stories, convene panels and feedback findings.
Outcome harvesting (Wilson-Grau, 2015; Wilson-Grau and Britt, 2012)	Participatory approach; stakeholders collect evidence of what has changed, then work backwards to determine whether and how an intervention contributed to these changes. Draws on IDIs and surveys.	• Provides retrospective learning about what was achieved and how, regardless of whether it was pre-defined. • Requires an understanding of when might anticipate change to have occurred. • Useful in context where relationship between cause and effect are not fully understood.	• Suitable when inputs, activities and outputs and the causal mechanisms between them are not fully understood as does not measure pre-determined outcomes. • Can identify unintended outcomes • Tailored to project and context; findings not generalizable. • Only outcomes informant aware of captured. • Resources intensive • Participation of those who influence outcome
Outcome mapping (Earl et al., 2001; Jones and Hearn, 2009; ODI, 2009; Research to Action, 2012; Smutylo, 2005)	Focuses on changes in behaviour, relationships, actions and activities of the people, groups and organisations it works with directly ("boundary partners") and how far changes contributed to outcomes. Consists of three stages: 1) intentional design – to establish consensus on intended changes; 2) outcome and performance monitoring – uses journals to chart changes in the indicators defined; 3) evaluation planning – helps the programme identify evaluation priorities and develop an evaluation plan.	• Used at the project outset to identify activities and the individuals, groups, organisations need to work with to realise intended outcomes.	• Process is more intensive because it requires meaningful participation from boundary partners. • Findings will be context-specific. • Participatory approach means individuals involved in the project gain an understanding of their role in ensuring programme is a success. • Challenges in participatory approaches of unequal power relationships.
Positive deviants (Andrews, 2015; Busza et al., 2017; Positive Deviance Initiative, 2017)	Explores an individual's or group's, behaviours or characteristics that have enabled them to succeed when the majority of peers have failed when faced with similar challenges, constraints etc. These cases can be identified by both participatory means and more quantitative modelling approaches.	• To discover the inputs and activities that have driven success and thus identify solutions that can be tested elsewhere.	• Small sample size • Reflects perspectives of individuals interviewed.
Process tracing (Barnett and Munslow, 2014; Davies et al., 2016; Talcott and Scholz,	Uses qualitative methods to determine relative weight of evidence for causal links between activities and outcomes. The evidence is used to confirm whether mechanisms match predicted hypothesis. Comes from the analysis of historical events.	• To see if results are consistent with the hypothesised mechanisms of action and to see if alternative explanations can be ruled out.	• Make strong causal claims about what mechanism(s) caused a given set of outcomes in any given case.

Method	Description	Use and timing	Strengths/Weaknesses/ Considerations
2015; White and Phillips, 2012)		• Intervention needs to be at a relatively mature stage and some level of meaningful change has occurred.	• Requires sufficient time and human resources to enable participatory iterations of analysis and discussion with stakeholders.
Qualitative comparative analysis (QCA) (Befani, 2013; Davies, 2016a, 2016b; Jordan et al., 2011; Kane et al., 2014; White and Phillips, 2012)	A theory-driven approach used to examine the relationship of *a priori* outcomes of interests and the conditions hypothesised to influence the outcome. Qualitative data is converted to quantitative data (either binary or ordinal data) and tabulated for each condition and outcome. Patterns in the resulting data table are identified to highlight pathways of conditions that produce an outcome.	• To test existing theories and new assumptions and formulate new theories. To understand the context under which interventions work and how different implementation strategies effect outcomes. • Potential to support short cycle learning about the effectiveness of specific activities being implemented during a project's lifespan. However, quite a time consuming process	• Provides causal inference. • Does not account for temporality. • Can use relatively small and simple data sets. • Strong external validity. Allows for the generalisation of findings from a relatively small number of cases and offers the ability to identify different pathways of condition combinations that lead to a similar outcome. • Do not need to pre-specify causal pathways between activities and outcomes. • May require more data as likely there will be a wider range of interventions and outcomes where relationships are possible.
Qualitative impact assessment protocol (QuIP) (Copestake, 2014; Copestake et al., 2018b, 2018a; Copestake and Remnant, 2015)	Outcomes are explored with programme or intervention intended beneficiaries, to identify those factors beneficiaries perceive to be driving changes. Interviewers are blinded to the theory of change and project being assessed. Ask about casual drivers of change in selected areas of respondent's life. Data is coded quantitatively, highlighting whether reasons given for change confirm the hypothesised causal pathways. Code whether evidence is explicit (i.e. referenced project) or implicit.	• Undertaken at a single point in time; although could be repeated to examine change over time • Particularly useful where evaluation has not been incorporated from a programme's outset or where the context is highly changeable. • Examines whether interventions having planned impact on intended beneficiaries. • Provides both confirmatory (e.g. to test theory of change) and exploratory learning (e.g. open to unanticipated drivers and outcomes).	• Does not require a baseline or comparison group. • Does not provide an estimate of magnitude of effect. • Quantitative coding of qualitative data speeds up data analysis. • Findings presented in a dashboard, make them easy to interpret. • Can identify unintended consequences • Reflects perspectives of individuals interviewed; may not be generalisable to other settings. • The QuIP incorporates features of a range of other qualitative approaches, including contribution analysis, process tracing, outcome harvesting and realist evaluation • Aims to addresses the challenges of confirmation bias through "blindfolding"

Method	Description	Use and timing	Strengths/Weaknesses/ Considerations
			• interviewers and respondents from knowing the full details of the intervention evaluated •
Rapid assessment process/Rapid assessment methodology (Beebe, 2001; Butler, 1995; Harris et al., 1997; Hildebrand, 1981; Manderson and Aaby, 1992; Schünemann, 2015; Vlassoff and Tanner, 1992)	Highly focussed team based ethnographic approach; uses IDIs, FGDs and observations. Three major features: 1) a systems approach; 2) triangulation of data; 3) interactive data collection process to quickly develop a preliminary understanding of a situation from the insider's perspective.	• Could be undertaken at any stage of the programme. • Undertaken at a single point in time. • Aims to collect only relevant and necessary data; makes more rapid and cost-effectives than traditional qualitative approaches. • Teams of interviewers may be used to rapidly collect information with the study completion expected within four to six weeks.	• Ability to adjust investigations to reflect local conditions and specific situations. • Involve the community in both defining community needs and seeking possible solutions. • Adopts the principle of adequacy rather than scientific perfection. • Subject to both respondent (courtesy bias, social acceptability/political correctness bias, positional bias/attribution bias, self-serving bias and self-importance bias) and evaluator biases (contract renewal bias, friendship bias, and similar-person bias).
Rapid review/Expedited review, Rapid evidence summary (Ganann et al., 2010; Grant and Booth, 2009; HEARD Project, 2018; Tricco et al., 2017, 2015)	A form of evidence synthesis. Methods vary; follows systematic review approach but places greater number of restrictions; e.g. fewer databases searched, time and setting restrictions or omits some processes to produce information in a timely manner.	• To identify new or emerging evidence on a topic, to assess what is already known about an intervention.	• Provides more timely information than a systematic review by omitting stages of the systematic review process. • Less rigorous than a systematic review; search is not as comprehensive, may not double screen/extract, limited interpretation of findings etc.
Root cause analysis (Hubbard, 2010; Peerally et al., 2017)	A method of structured risk identification and management. Not a single technique; a range of approaches and tools drawn from fields including human factors and safety science used to establish how and why an incident occurred in an attempt to identify how it, and similar problems, might be prevented from happening again. Analysis aims to establish a sequence of events to understand the relationships between contributory factors, the root cause and the defined problem.	• Typically undertaken to identify the cause after an adverse event has happened. • Can be used to forecast or predict	• Assumes linear causal pathways. • Findings will be context specific.

Method	Description	Use and timing	Strengths/Weaknesses/ Considerations
	Undertaken by a small team of stakeholders and facilitated by an expert.		
Statistical process control (Benneyan et al., 2003; Fereday, 2015)	Combines time series analysis methods with graphical presentation of data. Output or outcome data are plotted over time against statistical limits to identify if observed variation in an outcome deviates from the expected level of variations. Signals when the data deviates from predictions.	• To determine whether changes in processes are making a difference to outcomes and/or to detect potential differences arising from different implementation strategies between sites. • Undertaken continually throughout programme using data collected at standard intervals.	• Measures short-term outcomes. Limited measurement of longer-term impact. • Requires ongoing data collection. • Requires data collection, analysis and feedback to be completed as close to real time as possible • Able to detect process changes and trends from an early stage in the programme; does not rely on reaching a pre-specified sample size - data limits adjusted when reason to believe current limits are not appropriate to provide adequate signals for action. • Can change indicators or incorporate new outcomes overtime • Potential bias from temporal trends
Strategy testing (Ladner, 2015)	Participatory process for adapting theory of change over time. Initial theory of change represents best guess, which is examined on a regular basis to determine whether the assumptions are still valid.	• To articulate and capture changes in the programme theory. • A structured conversation undertaken with relevant stakeholders every 3 to 4 months throughout project.	• Participants must be willing to engage in an honest and reflexive discussion. • Findings will be context specific.

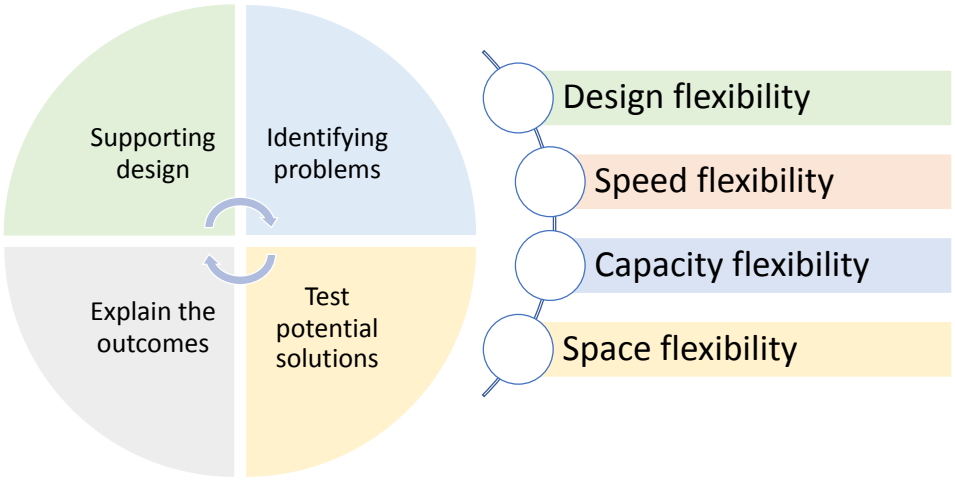

Figure 1: Framework for a timely approach to evaluation

Table 4 Appraisal of purpose categories of evaluation methods for application in timely evaluation

Method	Purpose				Dimensions of timeliness and flexibility			
	Support design	Identify problems	Test solutions	Explain outcomes	Design	Speed	Capacity	Space
A/B testing	Primary		Secondary		Moderate; test a priori outcomes - outcomes can be changed for each cycle of testing based on emerging information.	Rapid; measures short-term outcomes. Depends on timeliness of routine/programme data.	Requires statistical expertise, but potentially programmes/packages could be developed to be operated with less expertise	Potentially adaptive to changing contexts.
Adaptive RCT	Secondary	Secondary	Primary		Limited; modifications are pre-planned before data is analysed based on pre-determined decision rules. Outcomes specified at outset are maintained throughout.	Moderate; Combine phases of a trial, reducing the time needed.	High level design and analysis expertise required	Not adaptive to changing contexts once started. Decision rules have to be pre-determined.
Bottle neck analysis		Primary			Moderate; causal pathways hypothesised prior to data collection. Does not capture unexpected/unanticipated outcomes. If analysis repeated causal pathways can be adapted to reflect changes to the programme.	Moderate; depends on timeliness of routine/programme data. Slower if collecting primary data.	Moderate analytical expertise required	Can be adapted to include new/different hypothesised causal pathways
Modified stepped wedge trial	Secondary	Secondary	Primary		Limited; time built in between implementing steps to identify need for and make modifications	Moderate; implementation over defined phases of time	High level design and analysis expertise required	Not adaptive to changing contexts once started. Decision rules

								have to be pre-determined.
MSC		Secondary		Primary	High; can capture unexpected or unintended consequences	Slow; Takes time to collect stories, manage selection panels and feedback findings.	Interviewees can be trained relatively rapidly, and skills increases over time. Coding, analysis and interpretation requires skills & experience. Understanding of theory required.	Highly adaptive to different and changing contexts
QuIP		Secondary		Primary	High; can capture unexpected or unintended consequences	Depends on time taken to collect data. Reduced analysis time by converting qualitative data into quantitative.	Interviewees can be trained relatively rapidly, but skill increases over time. Coding, analysis and interpretation requires skills & experience. Understanding of theory required.	Adaptive to different and changing contexts
RAP	Primary		Secondary	Secondary	High; grounded theory analysis allows shift in focus based on emerging findings. Inductive adaptation of interview guides	Rapid; estimated to be completed in 5-6 weeks.	Interviewees can be trained relatively rapidly, but skill increases over time. Coding, analysis and interpretation requires skills & experience. Understanding of theory required.	Highly adaptive to different and changing contexts
SPC		Primary	Secondary		Moderate; decision rules apply, however, outcomes can be expanded &/or adjusted over time.	Rapid: Does not rely on reaching pre-specified sample size so able to detect changes from an early stage. Depends on time to collect, analyse and	Visual output is easy to interpret. Computer programmes available to support analysis.	Analyses highly adaptable across contexts, data recorded potentially difficult to change to include new indicators and data points recorded

						feedback data. Can be very rapid where routine data is readily available.		

TABLE NOTE: Primary = main focus of the approach; secondary = possible, but not a main focus