Designing evaluations to provide evidence to inform action in new settings

Calum Davey[1]

Syreen Hassan[1]

Nancy Cartwright[2]

Macartan Humphreys[3]

Edoardo Masset[4]

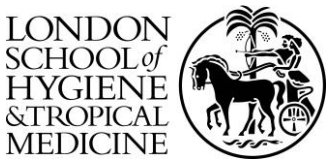
Audrey Prost[1]

David Gough[5]

Sandy Oliver[6]

Chris Bonell[1]

James Hargreaves[1]



[1] London school of Hygiene and Tropical Medicine
[2] Durham University
[3] Columbia University
[4] Centre of Excellence for Development, Impact and Learning
[5] EPPI Centre, University College London

Suggested citation: Davey C, Hargreaves J, Hassan S, Cartwright N, Humphreys M, Masset E, Prost A, Gough D, Oliver S, Bonell C, 2018 Designing Evaluations to Provide Evidence to Inform Action in New Settings, CEDIL Inception Paper No 2: London

About CEDIL: The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by UKAID through DFID. The mission of the centre is to develop and promote new impact evaluation methods in international development.

Corresponding Author: Calum Davey, email: calum.davey@lshtm.ac.uk

Table of Contents

Executive summary

Policy and interventions should be informed by the best available evidence, but evaluations are not always optimally designed to inform decisions about policies and interventions in new contexts. Learning the most possible from evaluations is important; evaluating is expensive and policy makers should be confident about their decisions. Using evidence from previous studies can lead to better policy decisions, but there have been cases where doing so has led to interventions that have not worked. Learning from evaluations for decisions elsewhere has generally been more successful for interventions that are simple and are less context dependent (or context-dependent in a simple way, such as depending on the severity of the problem). With increasing focus on complex, context-dependent interventions, we need to ensure that evaluations can offer as much information as possible to guide decisions in other contexts.

Consultation with DIFD to inform this paper underscored the points above. Examples where DIFD wants to learn more include:

- What has been learned from the recent outbreak of Ebola in West Africa that could inform future outbreaks, outbreaks of other diseases, or more generally about how health promotion can be reconciled quickly with cultural norms and expectations (such as to attend funerals and lay hands on deceased relatives)?
- What can be learned from the peace-process in Northern Ireland that could be applicable in South Sudan?
- What can be learned across evaluations of programmes that use mobile phone technology to change behaviours, both for future mobile-based interventions but also as a platform for understanding how habits can be changed efficiently?
- Large-scale, multi-component initiatives to improve the education system in a single country — what can the evaluation say about efforts to improve educational outcomes in other countries, and for engaging with public/private organisational cultures to affect change?

The aim of this paper is to suggest possible ways to address the issue of learning more from evaluations and make recommendations for how CEDIL could advance this area in the programme of work. To achieve this aim, we conducted consultations with experts from a range of disciplines to identify key concepts and developed a framework for possible approaches. We summarised and contrasted the approaches and reflected on their potential to address DFID's needs.

Methods

We consulted with members of the Centre for Evaluation at the London School of Hygiene and Tropical Medicine on the issue of 'learning for next time' from evaluations, and the design decisions that can help. We invited CEDIL intellectual-leadership team members to a half-day consultation and to be co-authors of this report. The co-authors were asked to provide literature relevant to these issues in their fields and to present the key insights during the meeting. In parallel with the consultations we conducted a

scoping review of literature on methods for learning from evaluations. We summarised key concepts relevant to our aims, identify four broad methodological evaluation approaches and suggest some recommendations for CEDIL.

Key concepts

The idea of transferring an intervention intact from one context to another has been described as retaining the original 'fidelity of form', i.e. keeping the intervention content as consistent as possible. However, adaptation is often needed so interventions are feasible and applicable to new contexts. Although it is tempting to identify and maintain fidelity of the 'core' of the intervention while allowing other parts to adapt, it is not easy to make the distinction. Evaluators can instead focus on the function of the intervention (i.e. the way in which the intervention is intended to generate outcomes) so that preserving fidelity of its function rather than form is the primary concern. However, it can similarly be difficult to know for sure how an intervention functions and how to retain these functions in a new context.

Middle-range theories are useful ways of bridging between the findings of an evaluation delivered in one context and decisions that need to be made in other contexts. Middle-range theories are theories that use abstract concepts to describe how processes occur. They are more general than the specific theories of change that describe how the inputs associated with a particular intervention will lead to its intended outcomes, but they are not as high-level as grand theories such as Marx's theory of class stratification or Foucault's theory of governmentality, which will offer very limited analytic traction for informing specific interventions. Theories such as role conflict (that people have finite mental capacity to cope with demands), cognitive dissonance, sociological theories about hierarchies and group dynamics, theories about management, educational theories about the role of metacognition in learning, and behavioural economic theories have many of the characteristics of what we mean by middle-range theories. Programme theories will tend to draw-on multiple middle-range theories to describe how a particular set of inputs should lead to outcomes, ideally with reference to supporting contextual factors. Middle-range theories can help transport knowledge by proposing how causal processes will play-out in particular sets of conditions. Evaluations can inform the refinement of such theories by testing whether their predictions with regard to specific interventions in specific contexts are correct.

While middle-range theories are promising, there remains the challenge of knowing when and where they will be useful for predicting outcomes (and therefore for designing interventions). Mapping each new context in its entirety is too costly and potentially endless. Finding 'markers' that are indicative of the middle-range theories that would likely apply in the given context is one way of addressing this. This goes beyond identifying the factors that moderate or mediate the magnitude of a particular intervention-outcome pairing (such as the effect of the prevalence of a particular disease on the magnitude of a preventative intervention). However, we did not identify techniques for findings 'markers', and the feasibility of finding such markers is still being discussed. A more pragmatic approach is to focus on identifying middle-range theories that suggest how social mechanisms play-out differently in different contexts, testing hypotheses and strengthening or refining the theories using empirical evidence. It

should follow then that when contemplating how to achieve a particular outcome via an action at a particular level (e.g. individual, community or societal), decisions can be informed by the theories that engage with the appropriate level and are sufficiently evidence based. The different theories will guide which aspects of context to think about by allowing predictions as to the actions will lead to what consequences in the context of interest.

Four evaluation approaches to building middle-range theory

The conceptual literature described above strongly implies that improving middle-range theory can strengthen learning for elsewhere and the design of evidence-informed interventions. We identify four evaluation methods that show promise for helping to strengthen theory:

1. **Framing evaluation questions for testing hypotheses to build theories**. Evaluations are generally designed to test interventions and not to test hypotheses to build middle-range theories. Most evaluations are set up to test whether or not the intervention works, and how it could be improved. These are legitimate aims, but evaluations can also be used to test the claims of the theory of change (hypotheses), and thus improve the middle-range theory that has informed it. Theories are strengthened when hypotheses that follow from the theory are carefully tested and found to be correct. Theories might be refined where results from an evaluation are not as the middle-range theory predicted. However, this seldom happens in practice. Hypotheses are not often tested, and the design of the evaluation only allows exploration of the plausibility of the theory. Moving to a position where evaluations aimed to test hypotheses that strengthen or refine middle range theories would probably require redesign of interventions. There might need to be more proof of principle studies that test particular mechanisms in particular contexts. This might involve more factorial or multi-arm studies that could focus on discrete mechanisms, and more purposive sampling of divergent context within or between studies.
2. **Process evaluations and mixed methods**. Embedded mixed-methods process evaluations are now normal in public health in the UK, and increasingly common in other disciplines and countries. Process evaluations can examine processes of intervention delivery and receipt. They can also examine the mechanisms by which interventions generate outcomes, and the salient characteristics of the context with which the intervention interacted, for example using qualitative, mediator or moderator analyses. Process evaluations thus provide one means via which evaluations aiming to build theories might occur.
3. **Leverage heterogeneity to understand context**. Since uncovering the interaction between interventions and context is a key component for building middle-range theories on how social mechanisms play-out differently in different contexts and creating transportable knowledge, evaluations should be specifically designed with a range of contrasting contexts (in terms of both characteristics of time, place and population). Researchers have recommended that policy evaluators should work in contexts that represent extremes of context rather than the 'norm'. The ideas of 'pragmatism' encourages evaluators to allow implementation to be 'real life'. However, there is evidence that policy

makers prefer to learn about which interventions can work and under which conditions and combine this with separate lessons for delivering interventions that can be drawn from across multiple sectors of public policy.

4. **Leverage heterogeneity using case studies**. An element of most process evaluations, and a complement to working in heterogeneous contexts, is the development of methods for selecting and learning from case-studies. Case studies offer in-depth exploration of causal factors by detailed tracing of the processes affecting outcomes.

Interventions and policy decisions should be based on well conducted synthesis of available evidence. Traditional meta-analytical methods may be useful for developing theory. However, they may be limited in depth because rather than exploring how mechanisms vary across contexts to generate different outcomes they restrict analyses to searching for patterns whereby homogeneous interventions implemented in homogeneous contexts generating consistent outcomes. Methods from the causal inference literature may be useful for drawing together context-specific findings for different effects and causal relationships in a transparent way. Once theories are more developed, and the core causal elements supported by evidence, they can be useful for extrapolating outside of contexts represented in the evidence base. With well-developed behavioural models, economists have had some success with extrapolation and prediction, and considering macro-economic consequences of wider-scale policy change could help predict how interventions will have effects over scales that would be almost impossible to directly evaluate.

Recommendations

CEDIL should focus on a number of evaluations and systematic reviews as exemplars of evaluations contributing to theory building where mechanisms interact with context. These should be important in policy terms. The evaluations should start with a synthesis of relevant literature. The evaluations should encompass a range of contexts (place and population) so that they are of immediate policy relevance and to help inform theory refinement. We recommend the following specific activities:

Create a Gap Map of theories in development. A Gap Map for theories in development is a taxonomy of the main theories that are used as the foundation of intervention designs. The Gap Map should be developed as a preparatory step in the design of a set of real prospective evaluations.

Form Theory Hubs. CEDIL should actively engage interdisciplinarity by looking outside of development and public health to build hubs of experts with relevant theoretical backgrounds, for example from business schools or marketing firms. The hubs should be guided by prospective and on-going evaluations.

Implement theory-focused synthesis methods. CEDIL should identify 2-3 key theory questions about the implementation of contemporary interventions and commission systematic reviews of the available science to generate questions for planned evaluations.

Apply existing and novel methods to represent theory. In the context of planning and designing prospective evaluations, CEDIL should support the application and development of methods for the representation of middle range theory and for putting 'theory' into theories of change.

Pilot theory-based evaluation. CEDIL should support a suite of studies in substantive areas of important to contemporary international development policy that try out the ideas we have discussed in this report, as well as a process of reflection and guidance on best practices arising from this work.

Identify markers. A key challenge is quickly identifying when theories will apply and will be useful for making predictions about the effects of interventions. CEDIL should consider this issue in substantive evaluations and reflect on the challenges and possible solutions.

Section 1

Introduction

It is expensive to test things everywhere and impossible to always test interventions, especially as contexts will change over time. Therefore, using knowledge gathered from particular places, people or points in time to guide activity for other places, people or points in time is necessary for timely and informed policy-making. But guidance on how to produce evidence that supports this process is lacking. In this paper we suggest that evaluations in international development, including those undertaken or commissioned by CEDIL, could be conceptualised, designed, analysed and interpreted in ways that could help us learn more for future international development policy actions.

The use of evidence from certain places, people or time points to inform decisions elsewhere has been successful in the past. We can reflect on the successes of evidence-based medicine as an example. We have found, for example, that BCG vaccination can prevent the acquisition of TB in many people. The evidence supporting the decision to administer the vaccination widely came from many places, research participants and points in time. This success-story reveals some of the challenges also: the vaccination does not work for people over a certain age, or for those with existing immunity (where the vaccination can cause moderate to serious illness). We have learnt, therefore, to accept that there are limits to the contexts (people) where the vaccine is helpful. This simple example of vaccines -- so simple that it might seem irrelevant to complex interventions working with social systems -- shows how pervasive the importance of context can be. The importance of context has been discussed extensively in the literature on the implementation of complex social interventions (Moore et al., 2008)(Egan et al., 2009).

When looked at differently, this example also hints at another way of thinking about context and using knowledge. Rather than using evaluations to think about the vaccination effects merely in terms of whether the intervention works or not in a certain context, we can (and TB scientists do) use evaluations to develop more elaborated theory about how it is that people acquire TB, how our immune system responds to exposure to pathogens and how vaccines can modify these processes. Evaluations of the efficacy of the vaccine in different circumstances can be used to refine these theories. It is these theories that tells us when to expect that the vaccine will reduce the acquisition of TB. The usefulness of this theory does not stop there; this knowledge can be informative for preventing infectious diseases more widely.

The problem of how to design evaluations that provide the most useful information to inform future decision making is a common problem in international development. In consultation with DFID conducted for this paper, we quickly generated several relevant examples. Imagine a programme of evaluation research to be conducted in a single country on a handful of interventions to improve educational outcomes for primary-school children. How could this programme of research be most usefully designed to inform education policy and interventions in other countries? Take the example of the recent Ebola outbreak in West Africa. How might evaluations conducted during and since that crisis be designed such that they can optimally help guide how to respond to the next epidemic of an

emerging infectious disease? How might we learn from peacebuilding in Northern Ireland in such a way that can inform programme design in South Sudan? How might evaluations of the deployment of new technologies (say mobile telecommunications, or social media platforms) for specific indications in one setting, be oriented to provide more general learning about their use for other applications in other settings?

Responding to this wide range of problems, the overarching aim of this paper is to identify, categorise, describe and appraise concepts and approaches that can increase the extent to which evaluations provide learning to inform action in new settings.

Section 2

Objectives

The specific objectives of this paper were:

- To identify key concepts and approaches that can be used in the conception, design and analysis of evaluation studies to optimise evidence that can inform action in new settings
- To provide a framework for categorising these concepts and approaches
- To define the concepts and describe the approaches we identified
- To reflect on the potential of these approaches to be useful for DFID in its evaluation activities
- To suggest areas where CEDIL's programme of work might extend this agenda further

It is worth saying one or two further things about our intended scope. First, this paper does not discuss trial or quasi experimental impact evaluation designs per se. There is a wide range of literature on characteristics of these designs which, first and foremost, are used to ensure that the results of evaluation studies have high internal validity. While we see internal validity as a necessary step towards drawing conclusions about actions in other settings, it is not sufficient, and our attention is oriented toward this next step. Second, we are firmly of the view that few, if any, decisions about policies or interventions should be made on the basis of single studies; rather, evidence informed decision-making necessarily requires synthesis of the evidence base. We see evaluation studies as part of a cycle that both starts and ends with synthesis, to inform the questions asked by evaluation studies and to update the knowledge base with the findings when they are complete. Since our aim was to explore how evaluation studies can most usefully inform future decisions, we necessarily considered concepts that are relevant to evidence synthesis and draw on these in our discussions. However, we have sought to stay principally focused on how primary evaluation studies can better provide information to support future decision making, through incorporating in synthesis, rather than on the synthesis methods per se. We reflect on some of the overlaps and implications of the approaches we outline for evidence synthesis in the Discussion.

Section 3

Methods

Consultations with experts

LSHTM Centre for Evaluation members: we invited members of the Centre for Evaluation at the London School of Hygiene and Tropical Medicine (LSHTM) to attend a morning consultation on the subject of learning more from evaluations. Around 50 people, from research degree students to distinguished professors attended the Consultation and participated in a discussion.

DFID staff: We had several communications with DFID, seeking to elicit examples of evaluation problems. We held one consultation with DFID's Senior Responsible Owner for CEDIL, in the form of a 1.5-hour meeting to discuss DFID's needs and examples of the areas where more could be done to support the transfer of knowledge between places.

CEDIL Intellectual-leadership team members: We invited the co-authors of this report, drawn from the CEDIL's Intellectual Leadership Team (ILT), to attend a half-day consultation at the LSHTM on November 21st, 2017. Prior to the consultation, the co-authors were asked to share key papers that offered important and relevant concepts and ideas. During the consultation, a drafted outline of the report was presented, discussed and amended. Key papers shared prior to the meeting were discussed. Scheduled "writing time intervals" were used by the co-authors to draft their ideas, opinions, and comments on the discussions and next steps were agreed. We used a "live" Google document so that several authors could write and comment on the text at the same time during the workshop.

Scoping review

In parallel with the Consultations, we sought to identify literature on concepts, methods and approaches for generating and using evidence collected in one set of people, places or time points to make inference to inform action in other settings. We sought papers in which any of the following terms appeared in the title or abstract: external validity, generalisable/generalisability, transfer/transferability, transport/transportability, application/applicability, adaptation/adaptability, knowledge/evidence transfer, transfer of knowledge/evidence. The terms were defined with enough precision to enable us to include papers that describe the concepts they defined even if the terms themselves were not used, for example, papers that discuss "whether knowledge from one setting is useful in another." Papers that did not use the specified terms or describe the concepts we were interested in were excluded. We also allowed modification of the inclusion criteria as the review progressed so that apparent gaps could be addressed and points of interest be pursued in more detail.

Our search strategy had three main stages. In stage one, authors identified important papers (key texts, widely regarded as theoretically/methodologically influential within the field) which they thought made important advances in the understanding of

knowledge transfer. Secondly, we screened the titles of the reference list of each of these suggested papers and subjectively appraised these for relevance. Third, we performed forward citation tracking to identify studies that had cited any of the studies included from first strategy. The ILT identified 57 key papers for review. From these, more than 200 references looked likely to be relevant to this topic; after screening the abstracts, 84 texts were included at stage 2.

We developed a tool for extracting the relevant information from the included papers, shown in Appendix 1. Stage-1 full-texts were reviewed, listed in Appendix 2, and 34 described relevant methods. We expected that due to our broad inclusion criteria some included papers would be those discussing the problem without offering explicit solutions. Therefore, the main feature of our tool was distinguishing papers that made methodological suggestions or demonstrated methodology through an applied example from those that only described the problem. We focused on transparency, practicality of application for different areas of development, and replicability. We reflected on the usefulness of the methods for informing the overall approach developed in consultation with the ILT. We developed a framework to describe the relationship of each method to the other.

Organisation of results

Many themes emerged from the consultations on how to learn more from evaluations to inform decisions. There was an emphasis on the breadth of research that can be useful for informing decisions. There was also an emergent consensus that much transferable knowledge comes in the form of better theories: theories about how relevant states (outcomes) arise in systems and about how interventions affect this, and that an important challenge to using theory is knowing when the mechanisms that are described in theories are likely to occur in new contexts. There was agreement about need for adapting evaluation designs so that these contribute more fully to testing/refining theories.

We describe key concepts under three broad headings, aligned with the overarching framework of the UK Medical Research Council process evaluation guidance for complex intervention. (Moore et al., 2008) First, we consider concepts pertaining principally to characteristics of interventions or programmes, and the likelihood that they would be implementable or have similar effects in other settings. Second, we consider concepts relevant to theories and models which can be used to help depict, understand or investigate whether phenomena and effects described in one setting have relevance in other settings. Third, we discuss context specifically, and focus particularly on the notion of “markers” that might be helpful in identifying salient aspects of context that can help signify if an intervention-outcome relationship may hold to a lesser or greater degree in two settings.

We then describe methodological approaches in the literature that are used to increase the utility of evidence gathered from an evaluation in one setting in informing actions in new settings. Our organisation of these approaches is intended to draw on the same underlying structure as the previous section. We first discuss how evaluations might be framed more explicitly to build theory rather than focusing principally on answering questions about whether this intervention worked here. Second, we turn our attention

specifically to process evaluation, and discuss how evaluations can capture data from multiple perspectives on characteristics of interventions, and how they are accessed and responded to by intended clients, that can help assess their likely feasibility and effect in future settings. Third, we discuss approaches to evaluation that are explicitly geared toward exploring effects across contexts.

Section 4

Results

Concepts

A number of distinct concepts were referred to in the literature and discussions concerning learning for elsewhere. These are listed in the table below with a short definition and summary of the contexts in which the concepts were typically found. These are used and discussed in more detail in the following sections.

Concept	**Definition**	**Context**
Fidelity of form	Extent to which form of the intervention (activities, materials, delivery) can be maintained in different settings	There is a tension between remaining completely faithful to how an intervention was previously delivered (fidelity of form) and making implementation feasible through adaptation
Feasibility of implementation	Extent of the barriers and facilitators for delivering an intervention in a new setting	
Adaptation	Changes to an intervention in a new setting	
Fidelity of function	Extent to which the function of an intervention can be maintained in different settings	Interventions can be described in terms of how they function within a complex system
Interventions as events in systems	Thinking of interventions as events and processes that interact with systems of people and institutions, instead of as having a fixed form or function	
Generalisability	Extent to which claims about a sample can be applied to a population	There are at least three different concepts used for extrapolating causal effects from an evaluation to other places, which differ primarily in the extent to
Transferability,	Extent to which claims made about one setting can be applied in another	

Transportability	(1) Extent to which an intervention can be moved from one place to another; (2) Extent to which causal inferences about one system can be used to make inferences in another	which a particular target context is considered.
Context-mechanism-outcome configurations	Theories about how context and intervention mechanisms interact to generate outcomes	Realist evaluation proposes that we look at what works for whom and why by investigating what mechanisms give rise to outcome and which contextual factors are needed
Middle-range theories	Theories that are general enough to travel but specific enough to be useful	Using evidence from one context in another may be facilitated with theory
Mathematical models of systems	Sets of related equations that aim to mimic patterns in data and underlying structures	Systems of equations are often used in economics and infectious disease modeling to fit characterisations of underlying causal structures to data
Markers of Context	What we can observe about a new context that tells us that a theory will likely apply	Mapping the entirety of any new context is not feasible and quickly identifiable markers could be useful

Intervention concepts

Fidelity of form, feasibility of implementation and adaptation

Most international development interventions comprise, implicitly or explicitly, a set of components intended to be delivered in some way. The concept of intervention fidelity (of form or design) relates to the extent that what is intended to be implemented is indeed delivered (Castro et al., 2004). In turn, this relates to intervention feasibility or applicability: the extent to which interventions can actually be implemented across settings (Burchett et al., 2013; Wang et al., 2006). Feasibility of implementation will differ from setting to setting depending on the structures and resources in place that support or act as barriers to delivery. Finally, a related concept is that of adaptation. Adaptation

of an intervention refers to intended changes made to an intervention when implementing in a new setting (Castro et al., 2004).

There is a dynamic tension between fidelity and adaptation because for evidence-based approaches being implemented in new settings, client characteristics, programme staff (and other resources), and administrative capacity may all differ from the settings where evidence was generated. This is particularly true for branded programmes that view consistency and comparability as a virtue. One response to this issue (Mihalic, 2002) suggests that when delivering an intervention in a new setting the key thing is to maintain optimal fidelity. Because we never have perfect information on which aspects of an intervention are its most important 'active ingredients', there is a risk that we may adapt out the active ingredients. Another response suggests that local adaptation is essential to ensure interventions are feasible and acceptable in the new setting (Dane and Schneider, 1998). A compromise position suggests that adaptation should be restricted to non-core elements (assuming these are known). These researchers point out that in the studies suggesting the importance of fidelity to replication of effects, none of the studies report on fidelity that is anywhere near 100% so it is unlikely that perfect fidelity is required (Durlak and DuPre, 2008).

Striking a balance between fidelity and adaptation does not guarantee that an intervention that has an effect in one setting will have the same effect in another, because in a new setting it may not trigger the mechanisms needed to produce the same effects. We address this issue in the next section.

Fidelity of function, and interventions as events in systems

A system-focused approach suggests that complex social interventions are best understood as 'events in systems' (Hawe et al., 2009). Consequently, interventions need to vary across settings because the systems differ. In this view, fidelity of function is more important than fidelity of form. Fidelity of function refers to the extent to which activities being delivered trigger the relevant mechanisms of action. Fidelity-of-function implies that it is the learning about how an intervention works that is the unit of knowledge-transfer from one place to another, rather than a particular effect size and a detailed description of intervention activities.

Consider the example of Frances Gardner and colleagues' work (2016) on parenting interventions and their effectiveness in different cultural contexts. Contrary to some expectations, this group found that programmes that were effective in high income settings were also effective in settings that were culturally very different. Interventions that had been developed in those settings were no more effective. The authors argue that this was because the underlying theory of the interventions was the same in each setting, regardless of adaptations of form. The authors defined comparable interventions according to the underlying theories about how they were supposed to work. Had they not thought about interventions in terms of function, interventions which on the surface appear different might not have been grouped in this way.

Aligned with these ideas, a 'pragmatic' view sees social systems as composed of networks of actors, whose habitual or instrumentally directed actions contribute to changes in outcomes of interest (Gross, 2009). An intervention (consisting of material

and non-material resources aiming to change actors' behaviours) may cause disruption that could potentially change the prevalence of the outcome. But, the intervention may also itself be affected, or even rejected, by the system into which it has been introduced. For example, in a trial of restorative practice (RP) in English schools evaluating whether this reduces bullying it was concluded that it was not 'the intervention' itself that reduced rates of bullying. Rather, the conclusions focused on how the non-material resources introduced into the school through the programme altered the way local actors behaved and thus transformed the school system (Bonell et al., 2017). Implementation of RP in schools might alter the way in which students and teachers interact with one another, not merely in relation to bullying, violence or discipline but in other ways that change the school culture. It is these changes that are important rather than the form of the intervention itself.

A potential problem with the concept of fidelity of function is that it might lead programmers to assume that once a way to trigger key mechanisms has been found then an intervention that worked in one place will work in another place. This may not be the case if the same problem in two contexts occurs for different reasons. For example, a peer-education intervention aiming to reduce HIV among men who have sex with men in 1990s USA (Kelly et al., 1997) was found not to have similar effects when delivered in Scotland some years later (Flowers et al., 2002). While issues of fidelity of form may be relevant, differences in the risk factors operating in each setting were also likely important. Whereas outreach education might have been effective in addressing lack of knowledge as a key risk factor in the USA in 1990, this risk factor may have been much less important in HIV risk several years later.

With conceptual parallels with the ideas of fidelity-as-function, Cartwright (forthcoming) contrasts between 'intervention-centred' and 'context-centred' approaches to thinking about what social policies are. She characterises intervention-centred approaches as those that focus on features of the policies themselves, while the context-centred approach starts from trying to understand the context where the policy is going to be used and the causal pathways to the outcomes that the context affords (similar to events in systems, above). She argues that even if we start with an intervention-centred approach, there will be implied cause-effect pairings along the causal pathway that are dependent on features of the context. Therefore, to understand how interventions will affect changes elsewhere we need to know that the structural features that afford the connections are sufficiently similar in both places. Considering context when describing interventions not only helps with predicting the effect of interventions elsewhere, it also helps with identifying new policies based on the causal pathways that are available in the context.

In the previous section we discussed the concepts of fidelity of form, adaptation, and feasibility. Those carry the notion that in an evaluation the primary 'object' about which we hope to learn for other settings is an intervention that is conceived in a certain way. When considering implementation in a new setting these concepts encourage us to consider how best to preserve their form, whether they would be feasible to deliver, and how they could be adapted. In this section we moved to consider interventions rather as events or disruptions interacting with the system to change existing levels of

outcomes (a context-centred approach). This encourages an emphasis on function rather than form, and has echoes in Realism which we will discuss below.

Transferability, generalisability, external validity and transportability

Various concepts have been used to describe the extent to which causal effects that are found in one place can be expected to hold elsewhere: these include generalisability, transferability, and external validity. The underlying idea behind each is to recognise that evidence of a relationship (in our case, the effect of an intervention on an outcome) gathered in one setting may, to a greater or lesser degree, have relevance in other settings.

'Generalisability' is used to describe the extent to which claims about a study sample can be applied to other populations, without specifying specific other contexts or groups (Burchett et al., 2011). For example, population sampling is used to recruit study populations representative of the population from which they were sampled. In evaluations of complex interventions we address similar issues, even if we may be less likely to imagine that there is an overall population effect that we are trying to estimate. Importantly, decision-makers in new settings will want to know what effects they can expect on their patch more than what worked 'on average'.

'Transferability' refers to the extent that effects observed in one setting are likely to be repeated in another setting, or target context (Burchett et al., 2011).

'Transport' is used in at least two ways in the literature. First it refers to moving a programme from one place to another, and is therefore similar to the everyday usage of 'transport' (Leijten et al., 2016). Second it refers to whether and how causal information about one system can be used to make inferences about another (Pearl and Bareinboim, 2014). The term therefore differs in scope from transferability; transport refers to the use of any causal knowledge and not only to whether or not the size of the effect is likely to be the same. For example, public-health practitioners may worry about when a TB vaccine transfers to a new setting, while other researchers may be more concerned with whether what is known about TB vaccination transports to vaccine development and delivery for different diseases. Or, having conducted a teacher-training RCT in a number of schools in a city we might wonder if the effect would be similar in another city (transfer), but we might also wonder if what we learned about training is applicable in general to developing interventions in rural settings (transport).

Mechanisms and theories

The concepts described in the previous section start broadly with the notion that in an evaluation the primary 'object' about which we hope to learn for other settings is a particular intervention, package of interventions, programme, policy etc. In this section we discuss concepts that relate less to whether effects associated with these objects are 'generalisable', but are more concerned with articulating, representing and testing whether theories that explain interactions within systems (mechanisms of action about relationships between inputs and outputs) can help us transport learnings from setting to setting.

Context-mechanism-outcome configurations

The philosophical concept of Realism suggests that the things that we have knowledge about are real, that they exist independent of our thoughts about them, and that these un-observable objects can influence one-another through real structures (Whitbeck and Bhaskar, 1977). Realist notions have led to the idea of 'realist evaluation' (Pawson and Tilley, 1997). Realist evaluators start from theories about how, for whom and under what conditions interventions will work, and use data collected in evaluations to examine how context and intervention mechanisms interact to generate outcomes -- what they refer to as Context-Mechanism-Outcome (CMO) configurations. Realist theories specify which mechanisms will generate which outcomes, and what contextual factors these mechanisms depend on. This approach to evaluation proposes that the effect of an intervention (or lack of) is due to the actions (or lack of) taken by actors in a specific context in response to the intervention, which trigger mechanisms that lead to outcomes. These CMO configurations are the explanations for 'what works, for whom, in what circumstances, in what respects, and how?'.

Middle-range theories

Throughout our consultation, literature review, and discussions among co-authors we have come back to the notion of "middle-range theories" as being a potentially valuable concept that might help us design evaluations that can generate learning that is useful for informing decisions in future settings. Across our discussions, we have seen a range of uses of this terminology. Over the coming few paragraphs we will describe this concept and convey why we feel it is potentially particularly valuable for this paper.

Middle-range theory been used to cover a variety of different kinds of scientific information, with the term 'middle-range' marking out that these theories are relatively abstract, can apply in different places or times but usually not always and everywhere, cannot generally be supported by any single study or handful of studies, and usually involve novel conceptualisation. A characteristic of 'middle-range' theories that we hope to harness is that they might fill the space between local descriptions about 'how this led to that in this context' and 'grand' theories, such as the laws of supply and demand. We illustrate the idea in Box 1 where we give a worked example using functional equations of a local theory that is implied by a middle-range theory which is itself implied by a grand theory.

In sociology, Merton describes middle-range theories as: 'theories intermediate to the minor working hypotheses evolved in abundance during the day-by-day routine of research, and the all-inclusive speculations comprising a master conceptual scheme' (Merton, 1967). An example of a 'master conceptual scheme' or 'grand theory' might be Giddens' structuration theory, which describes how social systems are reproduced through an interaction between structures and agents. An example of a middle-range theory might be 'role conflict': the idea that when incompatible demands are placed on people, complying with all of them will be difficult (for example, full-time work and full-time parenting). Role conflict theory is compatible with structuration theory, but it is more circumscribed and potentially more useful for informing action in different settings.

Box 1

Three levels of theory

Consider the following three theories over binary variables Y, X_1, X_2:

Theory 1 (grand theory): $Y = X_1X_2$

Theory 2: $Y = X_1$ in cases in which $X_2=1$

Theory 3: $X_1 = 1$ caused $Y = 1$ for individual i in context $X_2=1$

Here Theory 1, interpreted as a functional equation, provides a complete description of a simple process.

Theory 2 is implied by Theory 1 -- that is, if Theory 1 is true so is Theory 2, but the converse does not hold. Indeed different underlying theories could all justify Theory 2.

Theory 2 provides an incomplete description; here it is a conditional theory, which states how things work in a part of the domain covered by Theory 2. Alternatively a theory 2 of the form $Pr(Y = 1) = \tau X_1$, makes a claim about an average treatment effect which is implied by Theory 1 together with a belief about the distribution of X_2 (that $Pr(X_2 = 1)=\tau$).

Theory 3 is a specific claim about causal relations for an individual. Theory 3 claims no generality. It is implied by Theories 1 and 2, but does not imply either of these.

Within this set of theories Theory 2 is "middle-range"; it is implied by one theory and implies another. In a broader set of theories however Theory 1 might be midlevel, if for example it is implied by some Theory 0.

Something akin to middle-range theory is found in Jon Elster's work on 'explanatory mechanisms' (Elster, 2007). These are 'frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with indeterminate consequences.' Elster's basic elements of explanation are 'atomic' mechanisms - 'elementary psychological reactions that cannot be reduced to other mechanisms at the same level.' Atomic mechanisms, in turn, can be used as 'building blocks in more complex "molecular" mechanisms.' Examples of such mechanisms are: 'younger sibling syndrome', 'being pained by the thought that others think badly of one', 'a tendency to cooperate', 'a tendency to punish non-cooperators', and 'greater aversion to open displays of economic inequality than to hidden ones'.

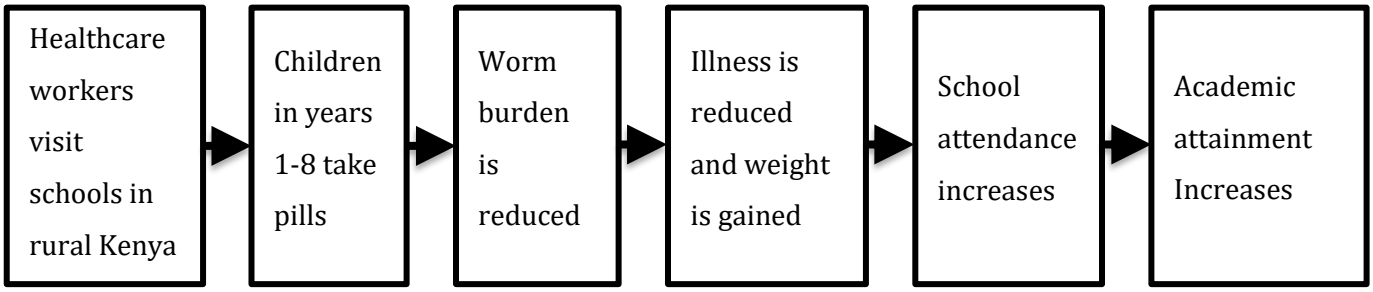
All theory uses 'abstract' concepts. However, some abstractions cannot inform practice until they are fleshed out in more detail. Thus, 'middle-range' theorisation usually involves the construction, investigation and validation of concepts that can be helpful to guide decision making. To give another example from economics, the grand theories that 'people respond to incentives' and that humans are 'rational agents who act to maximise their expected utility' are perhaps too abstract to be useful for making predictions in real situations. But, perhaps by fleshing these theories out with more concrete, 'middle-range' instantiations they can become helpful, and (critically for our purpose) perhaps transportable: for example, the middle-range theory that 'entrepreneurs looking towards their profits tend not to expand jobs if the available

workers have been unemployed for a long period and hence have lost skills' seems helpful, being both understandable in a given context but also challengeable or adaptable in others.

Middle-range theory shares features with the context-mechanism-outcome (CMO) configurations used in Realist evaluation, and the notion of 'explanations' from the inference-to-the-best explanation' perspective in the philosophy of science (Lipton, 2004). The idea of middle-range theory also resonates with the concept of fidelity as function in which interventions are seen in terms of how they interact with context to evoke changes in the desired outcomes. However, when focusing on interventions middle-range theories can sound very similar to the commonly understood concept of a 'theory of change'. Below we seek to clarify our thinking about some of the differences and overlaps in these ideas.

A programme theory, theory of change, or logic model is a description of how a particular set of activities is expected to set in motion particular causal pathways within a given context. The programme theory may include how the activities are expected to be implemented, or there may be a separate 'implementation' theory that completes the description of how the intervention is expected to work. For De Silva et al., a theory of change is a 'representation of the causal pathways through which an intervention is expected to achieve its impact within the constraints of the setting in which it is implemented' (De Silva et al., 2014). Here we use programme theory synonymously with theory of change, including the steps in implementation. Most programme theories are context-specific, use constructs that may not apply elsewhere, and link together too many different stages to be middle-range enough to be transported wholesale into a different context. However, although programme theory may usually be simultaneously too precise in its constructs and too unwieldy in how many elements are linked together to qualify as a middle-range theory, the constructs and causal assumptions within the programme theory may be local instances of middle-range theories. In particular, theories of change will almost always depict some causal process between programme and context element, most often as arrows but also as adjacent boxes, chains, or rows in tables. While theories of change are not middle-range theories, the arrows (or arrangement of boxes) and the rationale behind them, may represent where the authors believe middle-range theories apply. Theories of change are primarily used to communicate between stakeholders with very different backgrounds and should therefore be thought of as illustrations and not as scientific tools (for more on theories of change, see CEDIL report on *Representing Theories of Change*).

To illustrate, we consider a relatively simple intervention: delivering deworming pills to children in schools. A simple programme theory for this intervention may look something like the logic model below:



Our notion is that within this context specific programme theory, sit a series of local-level instances of middle-range theory that, if well understood, might transport to inform action in new settings. There is a middle-range theory about what influences whether pupils take pills and how and under what circumstances healthcare workers can influence this; a middle-range theory about what determines worm burden and how and under what circumstances taking pills can reduce this; a middle-range theory about child health and under what circumstances reduced worm burden can influence this, on what and over what time-frame; a middle-range theory about school attendance and how and under what circumstances health can have an effect; and a middle-range theory about attainment and how and under what circumstances attending school can help. An evaluation of this programme in a given setting can help us flesh out each of these theories, which if well understood could inform action across settings. Strengthening the theories further to inform action will draw on a range of sources, not only from evaluations, and certainly not only from evaluations of deworming pills. We return to the issue of synthesis in the Discussion.

Mathematical models of systems

Across the scientific disciplines represented through our consultations and literature reviews, it is a common practice to use mathematical models to emulate the real world. Such models are used for a range of purposes - to make predictions for the future, to synthesise and reconcile data relating to phenomena that are linked in non-obvious ways, and as a platform to undertake experiments that cannot be conducted in the real world. They seek to represent not everything that happens in a system, but only those characteristics of the real world that are relevant to understanding inputs and outcomes of interest to those in control of the model. As such they are representations of theory since the modeller prescribes the form of certain types of relationships within the model and determines what is considered relevant to include and what not.

In economics it is common to develop structural models of the system where an evaluation has taken place or where a new intervention is proposed (Deaton, 2009). Equations are used to describe the complex causal relations among the variables that matter in the site, but with sufficient construct flexibility to also, to some degree, apply in other places. An economics structural model often imposes a solution by assuming maximisation of utility or profits or social welfare. In this way shocks to the model have

a predictable outcome. Similarly, in epidemiology, infectious-disease models combine theory about how diseases spread, manifest, and cause illness, along with parameters from data, to describe and predict progression of disease in a population (Garnett et al., 2011). Like economic models, they combine assumptions about the structure of the process with data from different sources. They can represent current "best understanding", and by doing so they can include non-obvious things about their interactions. Mathematical models as they are used in infectious disease modelling rarely represent social or behavioural process, and so are more often used to think about questions such as: "what would happen to HIV incidence if we successfully treated STIs in this context", rather than those such as: "how can we incentivise the health system to treat STIs with high coverage". Indeed, structural models, both of economics and diseases, are currently most often applied to questions where there is not much debate about the extent of successful implementation. For example, they are focused on the effects of tax changes on purchase of alcohol or tobacco or the effects of STI-control on HIV-prevention. They can be used when compliance with the intervention can be accurately measured; they are less used where delivery is likely to vary, and as such the intervention is a less identifiable "thing".

Markers of Context

As discussed above, many conceptualisations of interventions see them as activities interacting with a set of already existing, and often changing, systems into which they are introduced. What final states emerge can depend on the set of interactions as a whole, or at least a good number of them. For prediction of what will happen in a new system, it is helpful to figure out which interactions (or mechanisms) were key when an intervention produced targeted results in a study setting, which are available in a new setting (both ones that can facilitate the success of the intervention and those that can impede it), what these more abstractly described mechanisms look like in the new setting, and what can call them into play at each stage. But we cannot map the whole system everywhere we go, and so it is helpful to seek contextual "markers" that can help guide us to knowing whether what happened here may also happen there.

Interventions operate in complex systems where they interact with many factors that might affect the final states that result when they are activated. For a set of actions to influence a set of outcomes in a given setting will require a set of "support factors" (sometime called "moderator" variables) to be in place.

Knowledge about what these support factors are is often less secure than what can be inferred from a well-conducted study on a given population but it can be exceedingly useful for predictions in new systems and in building new kinds of programmes geared to the new system. In some cases, we may have to know a great deal indeed about a new system to predict whether an intervention will initiate the expected sequence of changes there. But, sometimes there are ways to identify the right kinds of systems without studying their details, or even knowing much about what they need to be in order to support the changes hoped for. The philosopher Michael Strevens (2012) calls this 'Voodoo that works': you can sometimes point to systems with the "right" details without knowing what those details are. The trick is to find other kinds of information (other than the full system details themselves) -- what we are calling 'markers' -- that can

provide reasonable, or reasonable enough, indication of which theories (whether ToCs or other middle-range theories) might be trusted, where.

To help explore this idea from a different perspective, we can think of clinical diagnosis as being similar to the notion of identifying markers. In this case, diagnoses are middle-range theories; they describe how systems (bodies) give rise to outcomes (illnesses) and how various factors, including pathogens and treatments, interact. When presented with a patient who is short of breath, wasted, and pale, the doctor does not initially know why, and although she knows of many possible diagnoses (middle-range theories) she needs markers to indicate which theory applies in this particular ill patient. She elicits a patient history of symptoms (markers) and makes observations of signs (more markers). Ideally, this will be enough to make a diagnosis. For example, if the patient has been in contact with someone who has TB, and has symptoms such as persistent bloody coughing, we can start to see how just a few strong markers can tell her which theory probably applies. We can see from this example that more detailed diagnoses (theories), for example differentiating between sub-types of TB, may need more markers to know whether this theory applies here e.g. the results of microbiological tests. Identifying the cause of the illness with markers is, however, only the start: using the theory, she will need to consider the "support factors" implied by the theory that are needed for the treatments to alter the system to lead to desirable outcomes (e.g. patient age, weight, sex, comorbidities, other medications, allergies, values, preferences, and even ability to pay). If the patient does not respond to a treatment that is known to be highly effective this might act as a marker that the original diagnosis was incorrect.

For an example of how simple markers can be used in social settings, consider a systematic review of anti-bullying programmes (Lipsey, 2009). Researchers first developed a theory about how bullying is caused and manifests with distinct middle-range theories about how bullying works in younger children and older children. They predicted that a particular kind of programme would have different effects for younger and older children, using 'grade at school' to mark which of two complex systems of bullying children were part of. They found that programme effects did depend on age. Although these social-science theories are complex, we see parallels with the idea of diagnosis in clinical medicine, and with the idea of developing more specific diagnoses (here different kinds of bullying) as the first step in developing a response.

Consider a hypothetical example from political science. Two countries are in disagreement over a variety of issues, and tensions are mounting. Are they likely to go to war? The "theory of the democratic peace" or "inter-democracy non-aggression hypothesis" gives reason to answer *no* if they are both democracies. Roughly: democracies don't go to war with other democracies. To the extent that this middle-range theory is reliable, it can be useful in planning, not only military policy but also international investment policy. There are a variety of accounts of just what systemic features might be responsible for the democratic peace. But the point here is that understanding of the details of the systems that allow for democratic peace is not necessary for prediction: "democracy" is a relatively easily accessible marker for when non-aggression is likely. As with much in science, the theory is challenged, and much refinement has occurred over the years. Happily, not all the scientific issues matter for purposes of prediction. For instance, for prediction, in most cases it does not matter

whether the association is causal or merely a correlation. On the other hand, getting a good enough characterisation of when, for purposes of this theory, a country is and is not a democracy is essential if the theory is to provide a policy-useful marker.

Finally, for a development-centred example, consider what Deaton and Cartwright say about Conditional Cash Transfers (CCTs): '

> 'Conditional cash transfers have worked for a variety of different outcomes in different places… Think through the causal chain that is required for CCTs to be successful: People must like money, they must like (or do not object too much) to their children being educated and vaccinated, there must exist schools and clinics that are close enough and well enough staffed to do their job, and the government or agency that is running the scheme must care about the wellbeing of families and their children.' (Deaton and Cartwright, 2017)

Here many of the markers cited for identifying systems in which CCTs are likely to initiate the changes desired are akin to Elster-style mechanisms: individuals' desire for money, for their children to be educated, and for their children to be healthy, and the social tendency of a government to care about the welfare of its citizens. As is typical of Elster-style mechanisms, these are widespread but cannot be relied on to be universal: 'That such conditions hold in a wide range of (although certainly not all) countries makes it unsurprising that CCTs "work" in many replications, though they certainly will not work in places where the schools and clinics do not exist, e.g. Leroy et al. (Leroy et al., 2009), nor in places where people strongly oppose education or vaccination.' Many other potential barriers, such as policy-level attitudes to paying parents to educate their children, could also apply.

Four evaluation approaches to inform action in new settings

In this section we consider four broad methodological approaches to evaluations that offer the promise of explicitly providing more information that can inform action in new settings than is currently the norm.

Approach	**Summary**
Framing evaluations to test theories not interventions	Orienting evaluations to accumulate knowledge that refines theory, rather than testing the effects of interventions.
Integrated mixed-methods process evaluation	Gathering data on multiple elements: components of the intervention; implementation; mechanisms (mediators) and effect of context (moderators); representativeness of samples; risk factors; features of target place; practitioner experiences.
Leverage heterogeneity	Conducting multi-site and pragmatic trials to test assumptions in multiple contexts.
Leverage case studies	Using case studies to identify conditions where diverse outcomes are observed focusing on context, implementation differences, and "trajectories of change" to predict whether replication or scale up is possible.

Framing evaluation questions to test theories rather than interventions

Evaluations can be conceptualised as asking questions about theories rather than "Does this work here?" or focusing in particular on intervention effects. This way of thinking might be thought of as a natural extension of the conceptualisation of interventions in terms of their mechanisms (fidelity-as-function), or as events in systems (the context-centred approach). (This issue, and the degree to which different disciplines have focused on theory testing with evaluations in the past, has been discussed in another CEDIL report: *Epidemiology and Econometrics: Two Sides of the Same Coin or Different Currencies?*) Many development interventions are the result of complex negotiations between multiple parties who have different theories about what is important about, and what can be learnt from, an evaluation. As the realists attest: interventions will be informed by any number of theories, all of which are incomplete, middle-range, explicit or implicit. Interventions are "theories incarnate". We speculate that evaluations of complex interventions in international development could be more specifically oriented toward building middle-range theories that can be more useful for informing decisions in new settings than knowing whether this worked here.

Consider a large HIV treatment-as-prevention trial implementing a range of intervention strategies in health settings and communities. Any number of mid-range theories might be relevant to such a trial. Clinicians may see the critical step in intervention as about getting the right drugs to the right people and ensuring they are successfully treated. Epidemiologists may believe the critical aspect is to ensure that those who are able to get treatment quickly are those who most involved in transmission. Health systems researchers may see the intervention as principally a human resources experiment asking whether an entirely new cadre of health workers can be employed to deliver home-based testing, referral to treatment and adherence support. All of these conceptualisations are deploying middle-range theories, and yet all sit broadly within a single theory of change for the study. Writing down what an intervention is and is intended to do, and why, in the form of a programme theory or theory of change diagram is essential. But in addition, we might explicitly seek to reflect a range of middle-range theories that the designers think are relevant. Then evaluators can ask: what can we learn from this evaluation about the middle-range theories that are implicit in the intervention? For example, using the evaluation above: what can we learn about theories about how making health technologies more accessible can affect uptake? what can we learn about when and how interrupting transmission can control the epidemic? what can we learn about when and under what circumstances can new cadre of frontline workers be effective within an existing health system?

While we see promise in this idea, there will be challenges in operationalising this approach. Consider, a complex social intervention, a school-led intervention in English schools to reduce bullying through changes to the school environment with the promotion of social and emotional skills and restorative practices (Bonell et al., 2017). The evaluators hypothesised that "boundary erosion" was an effective mechanism to trigger reductions in violence particularly among working class students in working class schools, and that the key to boundary erosion was transforming relationships between students and teachers so that these were effective and not merely instrumental, with staff and students alike becoming better at understanding other people's perspectives.

They concluded that restorative practice and student involvement in school-level policy-making were effective ways to trigger these mechanisms. But some of these conclusions will be based on underpowered analysis, or those where confounding remains a significant issue. And how should others use these results? In the short term or in fairly similar settings other people might feel confident that they could achieve the same effects by implementing similar activities. But it may be more problematic if people in very different contexts tried to reduce violence by using other activities to erode boundaries. They would face considerable uncertainty about whether those activities were really going to trigger boundary erosion. In effect they wouldn't know whether their activities embodied fidelity of function. We speculate that the way round this problem is to orient evaluation and other research to accumulate knowledge in order to gradually refine theory (through synthesis) so that we have a much better store of insight about how to erode boundaries in different contexts.

A more radical suggestion from the literature is that evaluators go beyond asking questions about the theories relating to interventions to ask: why does the outcome occur at all? There is some evidence that decision makers want this kind of evidence: Burchett et al. found that decision-makers in Ghana valued research that increased understanding of local health problems over effectiveness evaluations because understanding the causes of outcomes was more important for choosing how to intervene (Burchett et al., 2015). Evaluations are an opportunity because randomisation can support researchers in estimating unbiased aetiological effects. Taking the example of HIV prevention for MSM in New York and Glasgow, learning that the lack of an effect in Glasgow was due to the underlying behaviours that give rise to risk was not only about the intervention, but informative about how behaviours lead to HIV risk in this population. Stephen Birch also noticed that decision-makers are mostly interested in how the health outcomes come about, and recommends a shift in the focus from providers and their services to populations and their problems (Birch, 1997). This shift in focus bears some similarity to the shift from being intervention-centred to being more context-centred that Cartwright (forthcoming) has suggested would help to improve social policy. This approach may also require a shift from linear 'chain' theories of change or logic models to a 'network' conceptualisation and representation that reflects the complex system. While evaluations may not always be able to use resources to help understand the causes of outcomes, by re-orienting to developing and testing the theories underpinning interventions there may be synergies that help with intervention planning.

Process evaluation with mixed methods

One framework for formally considering aspects of interventions within an evaluation, that is aimed at learning for new settings, is process evaluation. Process evaluation of complex interventions can now be considered mainstream within public health, and are common in a range of other sectors. While process evaluations span a range of aims, one is to help evaluators consider: 'If an intervention is effective in one context, what additional information does the policymaker need to be confident that another organisation (or set of professionals) will deliver it in the same way, and, if they do, will it produce the same outcomes in new contexts' (Moore et al., 2008).

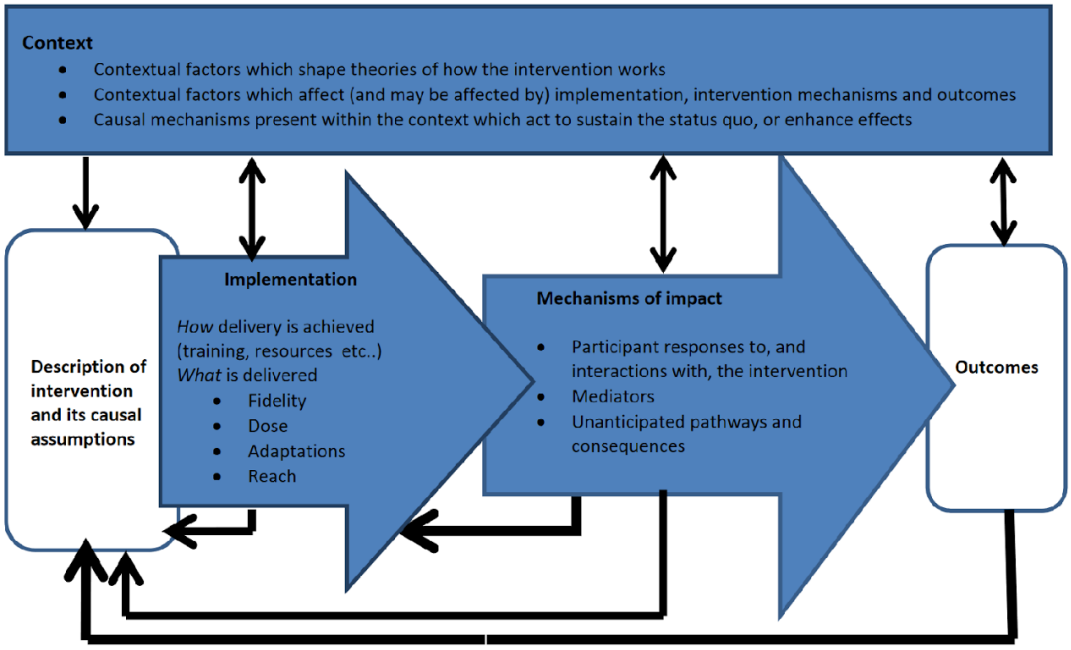
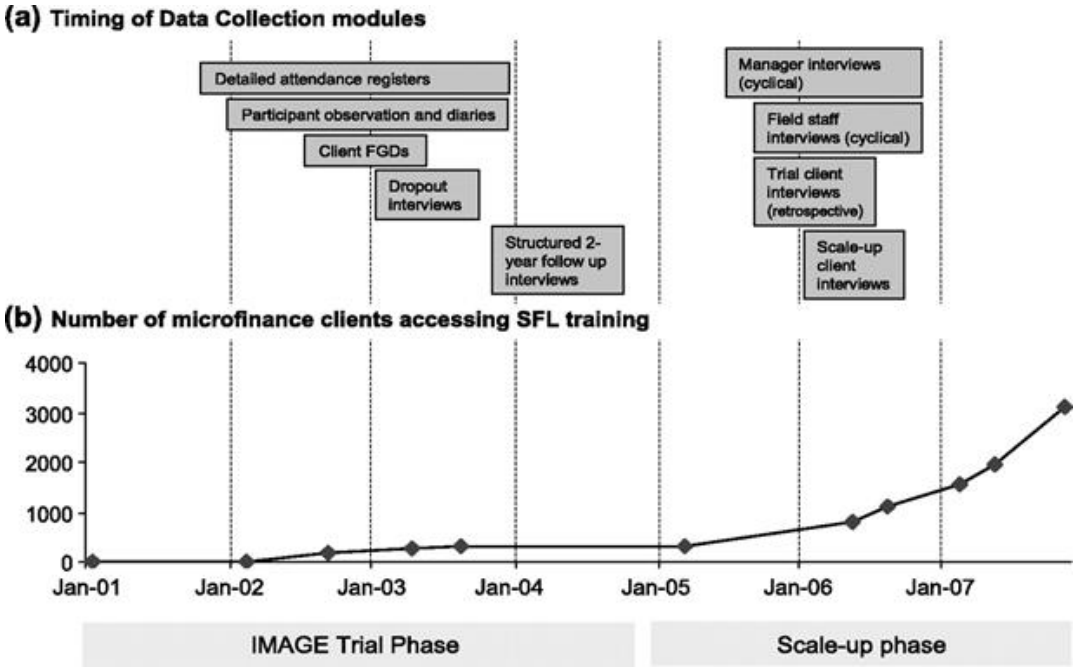


Figure 1. Key functions of process evaluation and relationships amongst them. Blue boxes represent components of process evaluation, which are informed by the causal assumptions of the intervention, and inform the interpretation of outcomes.

Similarly, Bonell et al. discuss generalisability in trial reporting and propose that researchers: integrate a process evaluation; develop evidence-based theories about how processes are affected by context and how they might differ in other sites; report the representativeness of participants in outcome evaluations; and, describe the needs being met by the intervention and its hypothesised mechanisms (Bonell et al., 2006). They acknowledge the usefulness of multi-site trials and highlight the importance of infrastructure and capacity in influencing how interventions are delivered and received, understanding how norms vary across settings which in turn affects whether interventions are acceptable to participants and other stakeholders, and assessing whether interventions are responding to needs (unaddressed risk factors) so that they might disrupt the processes that are generating the outcome to be prevented. In a similar fashion, Laura Leviton proposes combining methods to better assess external validity, including through better description of interventions and their contexts, and combining statistical tools and logic to draw inferences about potential future impacts. (Leviton, 2017) These approaches to process evaluation all require evaluators to develop sound programme theories which specify the core components of an intervention, deploy methods to help in documenting and understanding variations in implementation, and use a range of methods to identify the mediators (mechanisms) and moderators (contextual factors) of intervention effects. Leviton also stresses the importance of consultation with practitioners as their practical experiences can further specify core components and give information on the contextual features that are likely to be frequently encountered in practice, such as barriers, opportunities, and adaptations in response to these features.

Let us illustrate these ideas with two examples. Take the Intervention with Microfinance for AIDS and Gender Equity (IMAGE) programme in South Africa. This intervention sought to respond to an environment in which gender inequalities in economic opportunity and expectations were widespread, intimate partner violence and HIV infection rates were high, and communication about these issues was suppressed by pre-existing norms. The programme sought to bring together a group-based microfinance initiative aimed at the poorest women, with a range of discursive and participatory sessions ("Sisters for Life") on gender relations, cultural norms, sex, violence and HIV infection. The intervention was evaluated with a cluster-randomised, but also featured a nested multi-method process evaluation conducted both during and after the trial period (see Figure below for a schematic representation of the methods) (Pronyk et al., 2006).



The process evaluation had a range of aims, but let us focus here on those most pertinent to this section of the paper. The process evaluation collected: data from registers on whether women were able to attend the sessions they were intended to; structured questionnaire data on the acceptability of both the microfinance platform and the new gender-based education sessions to clients; informant interviews with clients, field staff and managers about the feasibility and experiences of implementing and taking part in the intervention package; and, a consultation with managers of microfinance programmes in other settings to assess the potential transferability of the model to other settings. Through this range of approaches the process evaluation sought to comment on the accessibility and acceptability of the programme in the trial settings, and its scalability and transferability to new settings (Hargreaves et al., 2011).

Process evaluations are also intended to help map data collection against the stages in the programme theory, theory of change or logic model. Such models, as we have

discussed above, are often constructed to be used in intervention or evaluation design or to guide the collection of data in an evaluation. Their focus is explaining how the activities involved in the intervention are expected to trigger certain mechanisms in a given context. We have discussed already how we believe a greater focus on middle-range theories might be particularly effective in supporting evaluations to be more useful in providing information to guide decisions in new settings. Process evaluations are one framework for considering the data needed to operationalise this approach.

Let us return to the example of deworming medication, to illustrate how process data might be used to improve the theories underlying the ToC. Imagine that evaluators persuade the implementers to deliver the pills at random. Whether "healthcare workers visit schools" is carried out by random assignment within a non-randomly selected group of schools in a non-randomly selected country. An outcome evaluation finds that pupils in the schools that healthcare workers visited had lower levels of worm infection, no difference in the degree of ill health (measured as anemia and anthropometry) and were more likely to attend school, but no more likely to do well on end-of-year exams. These effects have some claim to validity for this place and at this time, but they are also highly dependent on other factors: we would not, for example, expect the same effect on school attendance if there were dramatically fewer worms in the soil, or the same effect on worm burden if the pills were refused by more pupils.

To draw these conclusions data must be collected along the programme theory, and these data can then be used to test and refine theory. Take, for example, the middle-range theory that links the deworming programme to increased academic attainment, which would draw on research in education, psychology, and medicine. It may read something like this: "children who are fed, rested, and mentally and physically well, can perform well on age-appropriate academic tests when they attend a school with well-trained and motivated teachers who deliver an integrated curriculum that is relevant to the tests". Scrutiny of the process data from the trial may help test elements of the theory, using a mix of methods, to produce a better explanation of why children have particular academic attainment states. Consequently, the theory will be a useful starting point for developing interventions in other contexts, perhaps with a completely different configuration of interventions (perhaps even nothing to do with deworming).

Or take, as another example, the middle-range theory for attendance at school. It may have been, "children who are well will attend school when they and their families value schooling above the costs of attending: time, opportunity cost, and payments for travel, school fees, uniforms, books, and materials". The trial did not find any change in health but did see school attendance go up -- would we suppose that having worms should be added to the above theory? Perhaps before that, the evaluators may investigate whether the administration of the pills influenced the value that families place on school, or whether there were other parts of the intervention -- such as NGO branded materials -- which were distributed alongside. The theory directs these investigations, however open-ended methods might offer clues also: perhaps children who have worms feel embarrassed, or feel uncomfortable in other ways that were not captured by the measures of "health"? This could be investigated further; by reasoning towards the best explanation, the theory is strengthened. Process evaluations, underpinned by a theory of change against which data are mapped, are intended to ensure the data are

collected that can help in strengthening, weakening or adjusting programme theories that underlie the association between and intervention and outcome in a given setting, and thereby produce information that can guide decisions in other settings.

Leverage heterogeneity (i) to understand context

In a blog based on a talk given to DFID, researcher Chris Blattman, suggests that trials currently focus too much on "what works" and testing programmes, and that to address generalisability we should try to learn about the world more broadly by conducting multisite trials to explore local insights and test assumptions in multiple contexts (Blattman, 2016). To increase heterogeneity in terms of a particular contextual characteristic, Shadish, Cook and Campbell recommend choosing places that have the most common level of whatever characteristic, and extreme levels, rather than focusing on the mean (Shadish et al., 2002). Others have suggested that we need a more strategic approach to selecting interventions and contexts for evaluations, whereby we set out to test very similar or perhaps purposively slightly divergent interventions in purposively defined arrays of context in order to accumulate more systematic knowledge about context-intervention mechanism-outcome configurations (Bonell et al., 2012). Case-study methodology also offers good advice on choosing study sites, especially to matching the choices better with the kinds of questions asked and the dominant epistemic worries (e.g., more concerns over type 1 than over type 2 error).

A similar idea comes from pragmatism in the clinical and public-health trials literature that means allowing interventions to be delivered in "real life" conditions to see if the intervention can "work" (Schwartz and Lellouch, 1967). They are contrasted with explanatory trials, or efficacy trials, which test whether or not something can affect changes under ideal conditions. An explanatory trial might, for example, show that antiretroviral drugs for HIV reduce infectiousness when taken consistently, while a pragmatic trial may ask whether the same drugs can reduce the number of new infections when people take them with locally-appropriate levels of support.

The advantage of pragmatic evaluations of interventions for informing decisions is that they allow implementation and contexts to vary, and that they reflect experiences in intervention implementation in at least one "real life" setting. Decision makers may find that the places where pragmatic trials are delivered are more similar to their contexts than for the highly controlled intervention-contexts engineered for explanatory trials. Being pragmatic allows evaluators to explore and learn about implementation. Better information about the process of implementation, be that for managing and motivating staff to distributing resources, can be useful for decision-makers in its own right for introducing interventions. In interviews with Ghanaian decisions makers, Burchett et al. found that the impact evaluation results from effectiveness (pragmatic) trials were considered of less use than information about the efficacy of the elements of the interventions, combined with information about feasibility and adaptation, and evidence that the interventions would address the causes of the outcomes of interest in their context (Burchett et al., 2015).

Despite these advantages, even the most pragmatic evaluation is still just a "voucher" for effectiveness in other contexts (Hey, 2014). Pragmatic evaluations are, by definition, context specific (even when multisite), with the intervention model adapted for a

particular population and a context-specific comparison condition (the “control” group in a pragmatic trial will be highly context-dependent). There will still be differences in small and large ways from the wider range of target contexts and therefore however pragmatic the trial, it only provides indirect evidential support. Additional tools and techniques will be needed to interpret the result to inform activity in other places.

Leverage heterogeneity (ii) using case studies

Case studies have been proposed as a useful way to learn more from evaluations. By focusing in-depth in a few different locations, case studies may offer a cost-effective option for exploring how complex interventions are delivered. Woolcock proposes an analytical framework for using case-studies to identify conditions where diverse outcomes are observed focusing on context, implementation differences, and 'trajectories of change' (Woolcock, 2013). The three domains that are focused-on seek to engage with the complexity of the interventions, and case-studies are used to investigate the domains. The approach suggests first to identify what he calls the “causal density” of the intervention, that is how complex is the intervention and how many different pathways are being acted upon. Evaluators should then explore implementation capability required for the intervention and available in other contexts. This will lead to reasoned expectations about the potential impacts within the context, and when they are likely to happen. The focus of this framework is on whether something can be “replicated” or “scaled up”, and less on using the study to learn about systems themselves in the form of middle-range theory. The “complexity” that is addressed in this framework is more about the intervention than the context

In a recent paper, Cartwright describes a range of types of evidence that can be used to make causal claims in single cases (Cartwright, 2017). Some of these are similar to the Bradford Hill criteria for assessing causality (Hill, 1965). Other suggestions are to look for other symptoms of causation, such as side-effects, for the presence of support factors that are implied by theory, and to explore theory-predicted mediation through intermediate variables. These models align with our idea about middle-range theories and realist evaluation. The approaches that are discussed could be useful, alongside other small-n approaches (White and Phillips, 2012) for strengthening the face validity of case studies and increasing their influence on the interpretation of evaluations.

Section 5

Six possible ways forward

In this paper we have mapped concepts and approaches that might help us design evaluation studies that “learn more” than is currently the case. Too often, evaluation studies focus on identifying whether a particular intervention worked to produce desired effects in a given setting. Increasingly, it is common practice across disciplines to also ask evaluations to assess how interventions work, for whom they work and in what contexts. Yet there remains much to do to realise the ambition of an efficient evaluation agenda that is oriented toward maximally producing learning that can be

used by policy makers and programmers faced with an almost infinite constellation of problems and contexts. We have reviewed concepts relating to characteristics of interventions, programmes and policies, the representation and use of theory, and context characterisation that might be helpful in designing evaluations that learn more to inform decisions in new settings. And we have reviewed approaches to framing evaluation questions, studying the implementation of interventions in situ, and formally appraising contextual characteristics that can be deployed to this end. In this Discussion we circle back on three outstanding issues: what are the implications of this thinking for those who synthesise and use evidence to guide decisions, what are the potential benefits of further developing this agenda to evaluation as it is conducted in international development, and what further work should CEDIL undertake in this area.

Strengthening synthesis

In structuring the results of our consultations and review of the literature, we have focused on how primary evaluation studies can better provide information to support future decision making. But synthesis of the literature is crucial for informing the questions asked by evaluation studies, the results of which will update the knowledge base, which in turn will be drawn up for future intervention design and decision making. As we have seen, it is possible to learn more from an evaluation than simply "What effect did this intervention have here?". Synthesis methods have adapted to accommodate what can be learned about mechanisms and context, and also to provide ways of testing theories by using information from multiple sources. Thinking about how evidence can be used with synthesis can help us think about the usefulness of the various approaches that we have described.

Synthesis for interpolation

A number of synthesis methods use existing data to make best estimates of likely effects in new contexts by interpolating from the evidence that has been gathered in other places. By "interpolate", we mean that these methods do not try to apply prior evidence to places or interventions that are not represented in some way in the body of evidence.

Perhaps the simplest approach to synthesis is meta-analysis, which is often used in systematic reviews (Egger, Matthias, Smith, George Davey, Altman, 2001). It combines quantities -- usually the effect of interventions -- from different studies into a more precise overall estimate. Meta-analysing effects from different studies is reasonable where there is homogeneity of population and setting in terms such as individual, institutional or geographical descriptors, other factors that influence the risk of the outcome in question and the mechanisms that are triggered by the intervention to disrupt these factors to bring about a change in the outcome. In the example of MSM and risk of HIV, American MSM in the 1990s and Scottish MSM in early 21st century may appear to be similar but they were actually epidemiologically heterogeneous (Kelly et al., 1997; Flowers et al., 2002). In some cases, such as the review by Gardner et al of the transportability of parenting interventions to address children's conduct disorders, effect sizes appeared consistent across studies, suggesting that in these cases

populations were homogenous in terms of the important factors that support the intervention mechanisms (Gardner et al., 2016).

Variation in effects can be explored using "meta-regression" (Thompson and Higgins, 2002). This requires that evidence is collected in a sufficiently large number of different places, taking advantage of heterogeneity, and that the context or intervention elements have been adequately described in the reporting. Interrogation of heterogeneous effects should be theory-led so as to reduce the number of categories to investigate. Concepts such as intersectionality (for example, the life experience, health, and wellbeing of ethnic minority gay men in the UK are not just the sum of those of white gay men and straight black men) can, however, make identifying groups difficult to explore statistically using the available evidence.

Theory-led exploration of effects in meta-regression can be useful, however the focus is more on generalisability to non-specific places than addressing the transferability of intervention effects to a known context. Various authors have developed techniques that use information about the target context to weight the usefulness of evidence for making predictions there. The range of methods broadly falls into qualitative or quantitative approaches, and are briefly described below.

Quantitative re-weighting methods to explore transferability have been proposed in various fields including in public health (e.g. Susukida et al., 2017), and social sciences (e.g. Hotz et al., 2005). The need to re-weight the effect from the original studies comes from the observation that the effects of interventions often differ for different sub-groups, and differences between the original and target populations in the size of the sub-groups may mean that the average effect will not be the same. When estimating the average effects, the sub-group effects can be weighted according to the distribution in the target population. Examples of this approach are decomposition methods (such as, for example, the Blinder-Oaxaca decomposition approach (Fairlie, 2005)); predictions based on matching methods (Susukida et al., 2017); and sub-group analysis using machine learning methods for the definition of the groups (Athey and Imbens, 2015).

For re-weighting to be useful, information is needed about the evaluation and target contexts, and there must be overlap in the distributions of important variables. Larger-scale contextual factors can be more difficult to account for, since the amount of contextual variation on which data are available may be too small. While re-weighting can, like meta-analysis, be used to make predictions about the effect of interventions that are different configurations of components delivered in other places, it requires detailed information from the original studies about the exposures of individuals, or places, to different components, a convincing way to account for confounding (such as randomisation of different programme components, which is referred to as a "factor" trial), and also clear specification of the intervention components planned for the target population (Kelly et al., 1997). However, as it currently stands these data are not routinely made available from evaluations (Hoffmann et al., 2014; Hoffmann et al., 2013), and we hope this paper can contribute toward improving practice in this respect. Work is already underway to support the reporting and identification of important contextual factors, such as the TRANFER project at FHI Oslo (TRANSFER framework: Norwegian Institute of Public Health, 2017). As with meta-analysis, re-weighting on

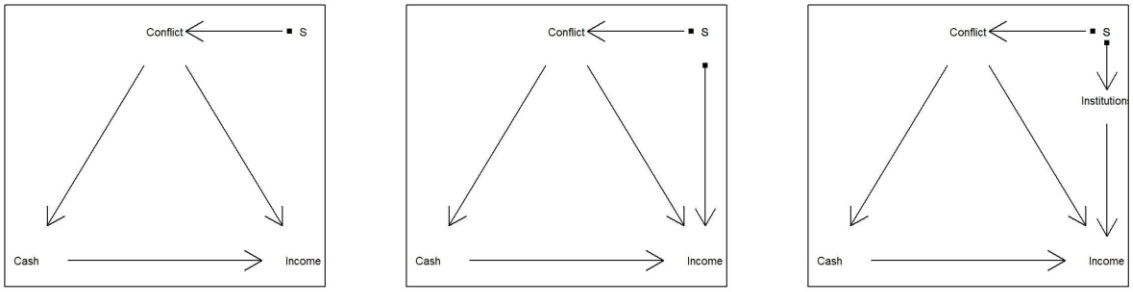
surface descriptors will not guarantee that the underlying structures have been accounted for. Re-weighting should be informed by theory, but often theory is not used to determine whether or not appropriate re-weighting is possible and what the re-weighting should look like.

To address the selection of variables to use when weighing evidence, Elias Bareinboim and Judea Pearl have developed a graphical procedure to decide if transportability is feasible and justify a transport formula based on theory (Bareinboim and Pearl, 2014; Bareinboim and Pearl, 2013).The broader term "transportability" applies here since the approach can be used in situations where there are a number of components (multiple observational and experimental results) to be considered jointly to make causal inference about a target context. Differences between populations are described using 'selection diagrams', and theory is used to define structural causal models (SCM). It is assumed that enough structural knowledge is available in both contexts to define the SCM. As they describe it, the algorithm 'takes as input a collection of selection diagrams with the corresponding experimental data from the corresponding domains, and it returns a transport formula whenever it is able to produce one.'

The core idea is that, where structural discrepancies between the two domains are suspected to take place, transportation requires controlling for the differences between systems so that the structure of the system, is no longer informative given the data employed in the analysis. To illustrate, imagine we are interested in the effects of cash transfers on income, knowing that conflict makes it less likely that individuals would receive cash and also less likely that they will generate income. Say we implement an RCT to assess the effects of case on income in a low conflict environment (e.g. Kenya) and want to make inferences for a high conflict environment (e.g. Somalia). We can consider three possible selection diagrams (see figures below), each representing our beliefs about processes in Kenya and Somalia and the differences between them. In the selection diagram in the left panel of the figure, the node S, representing a feature or features of the structure of the system, points only into Conflict. This means that the mapping from conflict and cash to income does not depend on S. If this graph is correct then it would be valid to simply re-weight the conflict-level-specific effects observed within Kenya to extrapolate to Somalia (i.e. using the re-weighting techniques described above, with caveats about the required overlap in distribution of whatever measure of conflict we are using). If on the other hand there were a second point of difference relating to other determinants of income, as in the second panel, then this re-weighting inference would not be justified. If, however, this additional difference could be represented as in the third panel --- as passing through an observable variable (this would count as what we have called a "marker" above)---here, institutions---then extrapolation again becomes valid, though this time by re-weighting the effects within strata that are formed by values of both conflict and institutions.

The advantage of this approach is that it can draw together information from various contexts to make inferences about the causal relations in a new context. It becomes possible to state justifications for transportation in terms of relations of independence between variables; and given a justification, there is clarity on the set of variables for which data should be gathered to calculate stratum-level causal effects.

The disadvantage is that it requires considerable confidence in one's understanding of the causal model in both populations in order to specify the key ways in which they differ. Constructing an SCM requires a minimal compatibility between graphs. For instance, the nodes are assumed to be identical in different contexts, which is a challenge for construct validity, and may require that latent variables are added to the model making it more complicated still. Despite these drawbacks, this approach offers a formal method for addressing transportability that is flexible, and further work could be conducted to apply these methods for evaluation and to develop tools to simplify and democratise the process.



Caption: Three selection diagrams. The first indicates that two sites differ in terms of the incidence of conflict; the second that they also differ in terms of the mapping from conflict and cash to income, and the third that the differences in this mapping passes through institutions, implying that conditional on levels of institutions and conflict the mapping from cash to income is the same in both sites.

A second approach to weighting evidence from other places uses -- broadly speaking -- more qualitative techniques. These are various tools that have been proposed to help appraise the applicability and the transferability of interventions for a particular context. They typically involve steps to collect data on interventions, the original and target contexts, and involve some degree of expert opinion. For example Wang et al (2005) suggested that a 'list of attributes that may impact on applicability and transferability can be developed, based on knowledge of the proposed intervention. Then the applicability and transferability of the intervention to the local setting can be rated, and given a score, based on knowledge of the local setting'(Wang et al., 2006). Burchett et al. reviewed available applicability and transferability frameworks and assessed how well they addressed settings, interventions, outcomes, and evidence (Burchett et al., 2011). None of the frameworks covered all four, and none were empirically based or tested. Most focused on appraising the applicability of the interventions and much less on support appraising transferability. Additionally, they found little published evidence that the frameworks had been useful, including for widely used frameworks such as RE-AIM (Gaglio et al., 2013). Further research could be done to test tools for assessing applicability and transferability in the field of international development.

Synthesis for extrapolation

As noted, there are lots of situations where re-weighting methods will not help with predicting the effects of new interventions, and we will struggle to account for large or unobserved differences in the contexts. This can be largely due to the reliance on interpolation between the existing data. To try to address this issue, researchers have developed more "structured" approaches so that they can make extrapolated claims. By "extrapolate", we mean that these sets of methods can be used to go beyond the range of contexts and interventions that are represented in the evidence. As with reweighting, researchers have developed both quantitative and more qualitative approaches to involving structure in the process of interpreting and using evidence to help with new interventions or very different contexts. Approaches include mathematical models of how people behave or how diseases spread (or both), and tools to support the adaptation of interventions to new places. By describing these approaches briefly here, we hope that we can show how they share conceptual links with middle-range theory and a context-centred view of intervention design.

As we have already described, quantitative, mathematical, structural models of systems are common in many disciplines. Economists will often approach a problem by first formulating a behavioural model of how people respond to policies (see, for example, Deaton, 2009) and Wolpin, (2013)). For example, we might start with a middle-range theory in the form of a structural model whereby students enroll in college based on a cost-benefit analysis of tuition fees and expected wages (Cameron and Taber, 2004)). Suppose that the researchers are able to estimate the effect of tuition fees on students' enrolment with some precision. Since the effect is operating through a change in expected costs and benefits, the modellers can use the estimated effect to predict the impact of other policies changing expected benefits and costs, such as, for example, an increase in wages. In this way, the impact on schooling of a policy that was never implemented, an increase in wages, but that shares the same function in the system, can be predicted. Thus, the economists' structural model uses a mathematical formulation of a middle-range theory to describe interventions in terms of fidelity of function, and therefore make predictions about interventions that share the same function if not the same form.

More recently, economists have used experiments in conjunction with structural models to test the behavioural assumptions made in their models (Schorfheide and Wolpin, 2016; Attanasio et al., 2011). In this approach, the researchers first develop a behavioural model, for example, of households' schooling choices. The model is then used to estimate effects of an intervention, for example responses to a cash transfer, using baseline data. The predictions are compared with the results of the evaluation; if the prediction was sufficiently accurate, then the researchers can become more confident to use the model more widely to assess the potential impact of policies that were not implemented, for example, for cash transfers of different sizes. This process bears many similarities with the idea of testing theory rather than interventions, recognising that a stronger theory may be more useful to policy makers than the particular effect estimated in that one evaluation.

Another approach to modelling the impact of policies consists of the use of general equilibrium modelling, for which a long tradition exists in economics. The rationale for the use of general equilibrium models can vary. In some cases, it is dictated by the scale of the intervention. Some interventions are implemented at a national or global level from the beginning like, for example, global environmental policies. In other cases, the interventions, for example an insurance product, can be implemented and assessed at the local level, but the results cannot be extrapolated at the national level, because they are bound to be different when implemented at scale. In a general equilibrium model, all the relevant behavioural responses, and all their interactions, are considered and any policy impact can be estimated. This idea from economics has some parallels with the idea that interventions are events in systems, and economic theory allows something to be said about how the system will respond.

These approaches are appealing as they allow answering a large number of relevant questions. There are, however, limitations. First, modelling behavioural responses or the behaviour of an entire economy requires great effort and skills and the availability of good data. Secondly, the results provided by these model simulations are not as reliably unconfounded as the results of an RCT. Finally, the results of the modelling exercises are as good as the assumptions used in their formulation. Economics models are, to some extent, middle-range theories, based on direct or historical observations of events, on social and psychological theories and observations or experiments. Economics models, however, share some fundamental assumptions about people's behaviours, namely the rationality and maximising behaviour of economic agents. These assumptions are at the same time a blessing, because they allow finding a "solution" to the models and predicting behaviours, and a curse, because the conclusions are valid "as if" people were behaving the way we expect, which is not always the case.

In public health, researchers have used infectious-disease models as means to aid the interpretation of RCTs, especially trials of complex interventions (Garnett et al., 2011). These models can include elements of context, and non-linear processes, and thus allow a better understanding of the effect size and ability to generalise to other contexts by linking trial outcomes to different intermediate and process measures, which can subsequently be compared with values from other settings to explore the potential effect of the intervention. For example, in HIV control, mathematical modelling helped uncover the generalisability (or lack thereof) of enhanced STI control as a method for controlling the HIV epidemic (White et al., 2004). Early RCTs gave different results as to how effective this was at reducing HIV incidence -- modeling showed that the key issue was most likely how prevalent and among whom curable STIs were present. While these models are useful for infectious diseases, and can be helpful during outbreaks (Funk et al., 2016; Rivers et al., 2014), they are developed and maintained by specialists, and draw so heavily on the basic epidemiology of diseases that they may not be directly useful in other fields.

Other modeling approaches have been used in political science. For example, Humphreys and Jacob propose a method to integrate quantitative and qualitative data to simultaneously update the strength of belief in a theory and produce estimates of

the effects of intervention in the context (Humphreys et al., 2015). The updating of the theory is made possible using a Bayesian framework that ascribes prior probabilities to the structures; observing the data can improve the precision of the probability bounds. This is a promising approach, but it requires that prior probabilities are ascertained, which may be too much to ask. However, tools could be developed to reduce the number of decisions that programme managers and evaluators need to address and to simplify the presentation of the method.

As we said, aside from mathematical models more qualitative tools have been developed to help decision makers adapt interventions delivered in other places to their context (Davidson et al., 2013). For example, building on processes for cultural adaptation, Castro et al. proposed a programme design strategy for "hybrid prevention programmes" that "build in" adaptation to enhance programme fit while maximising fidelity of intervention and programme effectiveness (Felipe González Castro et al., 2004). The approach identifies the core components while incorporating the values of the target population to refine these and develop new components. In another example, Wingood and DiClemente (2008) proposed a model to guide the adaptation of evidence based interventions in HIV prevention, called the "ADAPT-ITT model". This consists of eight phases for adapting interventions: 1) conducting focus groups with the target population to identify risk and behavioural context, and with staff to assess the capacity; 2) reviewing the available interventions and selecting which are appropriate for adoption or adaptation, based on composition of the target context; 3) adaptation using "theatre tests" with members of the target population; 4) drafting one of the adapted interventions, balancing need to retain fidelity to the core elements and underlying theory with the context's requirements; 5) consulting topical experts for advice; 6) integrating content from the topical experts into the intervention design; 7) training for delivery of the intervention; 8) testing by piloting with participants. Formative research is used to inform the selection of the interventions to adapt, and the theory of how the intervention is supposed to work is used to identify the "core elements". This approach implicitly accounts for the elements of the context that theories about behaviour change suggest are important and allows for the intervention to be adapted within bounds of the elements that are thought to contain the key mechanisms.

In an article titled, 'The Generalizability Puzzle', Bates and Glennerster describe a generalisability framework used at the Abdul Latif Jameel Poverty Action Lab (J-PAL)(Bates and Glennerster, 2017). The framework begins with the question, 'what is the disaggregated theory behind the program?', where 'theory' is described as ways of simplifying the world to help make and test predictions. The framework then focuses on the local conditions (and whether the theory is likely to apply), the evidence available to support the behaviour-change underpinning the intervention theory, and considers the evidence that the programme can actually be implemented in this setting. With reference to cases from various contexts, the authors argue that superficial contextual differences might not be important when mechanisms travel (such as educational interventions from India to the USA) and that learning about theory will require synthesis methods that interrogate the mechanisms behind the interventions. This framework bears many similarities to the overall approach that we have described.

In summary, synthesis is a key part of the evidence generation and use processes, and we think that in whatever form it can be better supported by a shift in focus to understanding the theories underlying interventions that are evaluated. Realist and theory-based reviews explicitly state this focus (Pawson et al., 2005), and they may offer a useful model to contrast with other methods. However, as we have shown in our brief review of example approaches above, many of the methods currently used to interpret and synthesise research evidence already have a role for theory.

How might this thinking help with DFID's examples?

At the start of this report, we identified some challenges, raised in consultation with DFID, to which the methods discussed in this paper are hoped to respond. We have argued that evaluations should be designed to test hypotheses and improve theory so that this refined theory can inform interventions and policies in other settings, which in turn generates new hypotheses to be tested in the new contexts to further improve the theories, and so on. This approach differs from a focus on conducting evaluations focusing on rigorous methods to establish whether or not an intervention works, without much theoretical grounding. Inspired by clinical trials units, evaluation professionals and organisations conduct randomised controlled trials in a wide range of disciplines to guard against subjectivity about the effects of interventions by letting the methods 'speak for themselves'. However, as we have shown in this report, although the methods are rigorous and do guard against subjectivity, without substantive theoretical knowledge it is difficult to transport findings from evaluated settings to other places. Since we are advocating for more theory in evaluations we cannot simply resort to describing methods in responding to the examples given by DFID of challenging evaluation contexts. Methods are used to answer questions and will require evaluation expertise, but the questions should be set by disciplinary experts working in the fields, for example, of psychology, sociology, behavioural economics, business, and marketing as applied to development. Therefore, in the examples below we have sketched illustrative responses that suggest how focusing on theory would change the evaluation, but we recognise the need for input from disciplinary experts to in practice.

How can we learn from one Ebola outbreak about how better to prepare for and respond to the next?

In an Ebola outbreak, an important factor that contributes to the spread of disease is norms about caring for ill family members and burial rituals (Manguvo and Mafuvadze, 2015). Convincing people to adopt new or uncommon behaviours is required to minimise transmission and control the outbreak, but this can be difficult. To better inform intervention designs elsewhere, it is possible that evaluations in such a setting could help refine middle-range theories about how norms and behaviours in times of crisis can be influenced, rather than principally studying whether or not an intervention introduced to change care and burial practices was successful in this instance.

Thus, we might focus on learning more about what kinds of messages can change social norms, rather than what specific messages do or do not work to change current practices. Possible middle-range theories could build on the diffusion of innovation theory, psychological theories of emotions and moral duties (e.g. Scott et al., 2007, and

Haidt, 2003), or evaluations of interventions that successfully altered norms with similar conflicts, such as those addressing genital mutilation.

For example, building on norm theory, which describes how descriptive (what others do) and prescriptive (what is appropriate) norms influence behaviour, researchers have hypothesised that a perception of an ongoing change results in a belief that these behaviours will become normative in the future and lowers the perceived barriers that prevent the behaviour, causing people to adopt those behaviours. They referred to this as "dynamic norms" to mean behaviours that are in the process of changing, and contrasted it to descriptive and prescriptive norms, which they described as "static norms" or behaviours that are already established. To test this, they varied messages to explore how people can be encouraged to eat fewer animal products (Sparkman and Walton, 2017). They designed messages that stated that people are eating less meat (creating the perception of a dynamic norm, or an ongoing increase in the prevalence of the behaviour) while other messages stated the number of people who now eat little meat (how norms currently are). They found that the former messages were more effective in reducing the consumption of meat, which supported the hypothesis and strengthened the theory of norms.

This approach could be applied to the context of an Ebola outbreak. We might predict that changing the norms around caring for the ill and burial practices could be achieved by delivering messages that portray "dynamic norms", implying that others are changing their behaviours, in various different contexts (e.g. urban and rural, or in different countries). Using process evaluation and case studies, the evaluation could collect data to explore how and why dynamic norms were or were not able to overcome the cognitive dissonance that arises from conflicts between promotion of behaviours that can help prevent serious illness, and cultural expectations that are required to maintain social status. Therefore, by observing whether, and how, the messages have effects, the theory would be strengthened and adapted to times of crisis. The evaluation may need to attend to how mechanisms vary, perhaps by conducting evaluations in purposively selected contexts. We might also do studies looking more generically at hygiene practices in family care and burial, not only in the context of crisis. Since the aim is to learn more about how to change behaviour in crisis contexts, we could do more basic studies looking at how norms based persuasive messages (not necessarily linked to Ebola or even infectious diseases) vary in their effectiveness with different populations in different crisis settings. A strengthened theory may be useful for designing responses to outbreaks of other diseases, earthquakes, floods, fires, and wars.

Clearly responses to Ebola outbreaks are highly complex and involve multiple actions, which might not be reflected in the above suggestions for strengthening the evaluation activity for learning more. We imagine that hypothesis-testing for building theories about changing normative behaviours in times of crisis would be just one part of a comprehensive evaluation that would capture both the traditional evaluation aims (did the intervention work? what happened?) and collect data to test other hypotheses implied by the middle-range theories behind other elements of the complex intervention at multiple levels, for example theories about human resource management, aid accountability, and political stability. Since evaluators would have little or no control over many elements of the response, especially the most critical

elements such as decisions about whether and where to deploy armed forces or enforce quarantine, evaluators may need to draw from across methods, e.g. between-place comparisons, case studies, interviews, to test hypotheses.

What can we learn from peacebuilding in Northern Ireland that can inform programmes in South Sudan?

In order to inform how evaluation design might be used to learn more by building theory, we might start by considering some of the theoretical work that has attempted to understand the Northern Irish peace-building case. For example, in a chapter 'Security is Not Enough: Ten Lessons for Conflict Resolution from Northern Ireland' in a report on the lessons learned from Northern Ireland, the Powell discusses lessons that can be learned from the Northern Ireland peace process (Powell, 2011). One of which is "[that] insurgent groups will not just surrender…[they] need a narrative to explain to their supporters what they have achieved and why all the sacrifice was worthwhile". This brings to mind the sunk cost fallacy, whereby decisions are influenced by the emotional investment that has been accumulated and the more investment the harder it becomes to abandon a project even when there are zero (or negative) marginal benefits to continuing to pursue it (an idea that is prominent in behavioural economics and in 'nudge'-style interventions). Insurgents may be less willing to make compromises as their investment (lives lost and sacrifices made) have accumulated. The report derives that for this reason it is a mistake to insist on preconditions before beginning talks (e.g. when governments refuse to be seen beginning talks with insurgent groups until there is a ceasefire) and argues that there should not be pre-conditions before talks can start.

Another lesson that was discussed is that peace is a process not an event, and that having a process in place means that people are hopeful and kept busy as violence may emerge when there is a void created by a lack of process. This lesson resonates well with the Behavioural Momentum theory (BMT)(Nevin and Shahan, 2011), which has been applied in education to increase student compliance to instructions by teachers(Oliver and Skinner, 2002). The analogy is the momentum of objects. Interventions based on the BMT administer quick requests with a high probability of compliance, and the response with compliance to these requests is re-enforced (e.g. with praise). This is referred to as rapid high probability request-response-reinforcement (RRR). Analogous to the momentum of objects, the high rate of reinforcement (i.e. force) increases resistance to change, thus increasing the probability of compliance when the low probability of compliance requests follows. We can see how this may apply to the lesson from Northern Ireland described above, whereby peace should be thought of as a process that can be further strengthened by designing to incorporate elements based on the BMT. Perhaps this also resonates with theory underpinning restorative practice, that conflict resolution requires us to see our adversaries as fellow human beings who can share their feelings with us, and with the sunk-cost fallacy; the more the parties invest in the process the less they are willing to walk away from it.

In the case of peacebuilding it is hard to imagine the scenario of a specific intervention/mechanism piloted in purposively varying contexts, or of an evaluation that tests these predictions using the classic rigorous methods (both because

randomisation is unlikely and there may not be that many contexts to compare). Selecting case studies to closely monitor when theory-informed approaches are applied, such as not placing pre-conditions before talks begin or designing negotiations to start with requests with a high probability of compliance first, followed by requests with low probability of compliance, and tracing the processes that led to the outcomes is particularly useful here. Context will be important: it may be that the long preparatory stage that involved back-channel (and therefore deniable) negotiations in Northern Ireland were a key element of the context in which sunk-cost considerations were made, and therefore this, and other aspects of context, would need to be considered when evaluating in a new setting. It might also be necessary to use non-direct evidence, for example rather than the key evaluations focusing on peace building they might involve lab studies looking at the mechanisms we are interested in but in lower stakes scenarios. To inform the South Sudan policy we might draw on studies of less serious and more common conflicts e.g. community disputes.

How can mobile telecommunication capacity and social media platforms be optimally leveraged to improve wellbeing and human development?

With the widespread use of mobile phones and social media platforms, it is reasonable to attempt to employ these channels to improve health and wellbeing in international development settings. The questions that should be asked when evaluating the use of these channels should go beyond what the best platform is, who should be targeted, or how many people engaged. For example, an evaluation could be designed to test theories about the relationships that people develop with these technologies, the types of messages that should be used, and how habits form. The importance of evaluations testing theories not just specific interventions is particularly clear here because any particular intervention will almost certainly become obsolete by the time that an evaluation, particularly if it is phased, is complete.

For example, in a study by Young, the author examined the effect of the sources of messages about health behaviours posted on online social networks by varying the messages that people received (Young, 2015). The predictions were derived from construal level theory of psychological distance that people report greater agreement with the health messages when they perceived the sources of those message to be from people of like attitudes and demographic characteristics. It would be possible to test the implications of this theory in different settings and for different development issues by designing an intervention where messages of different purposes, such as educational or invites for participation in a programme or activity, are delivered by those with or without similar characteristics to those of the target population. By contrasting evaluations of theoretically similar interventions in a number of settings and for different development outcomes, we would leverage heterogeneity to learn more about the role of context in determining the mechanisms behind effective telecommunication campaigns. Through this approach we would learn about the mechanisms underlying behaviour-change rather than tying our findings to particular technologies that, as we have said, are quickly likely to be obsolete.

A programme of evaluation research has been conducted in Ethiopia on a handful of interventions to improve educational outcomes for primary-school children. Improving

education is an aspiration for many DFID-focus countries. How can the programme of research in this one country be most useful to inform education policy and interventions in other countries?

DFID's project on improving the quality of general education in Ethiopia aims to improve learning conditions and improve capacity to enhance learning outcomes. Of the many aims, the school improvement programme aims to provide grants to schools to support their operating costs and create favourable learning environments for females and marginalised populations and allow for school-level planning of how these grants should be used.

Although not explicitly stated, this understanding the effects of this programme may be informed by Contingency Theory, which is one of many organisational theories that comes from the discipline of management studies. This theory says that there is no best way to lead a corporation or make decisions about its organisation, rather, the best way is dependent or 'contingent' on the leader's style of leadership and the specific situation in which the leader is acting.

It is possible that we could learn more from the evaluation of this programme by applying a Contingency Theory lens. The intervention narrative does not stop at providing schools with grants and allowing for school-level planning; the success of these components will depend on the planning decisions that will be made, and good planning decisions will be contingent on the leadership behaviours and the needs and environment that are specific to each school. An evaluation of this programme could take theory into account and assess the styles of leadership (task-motivated versus relationship-motivated), and how favourable or unfavourable a situation is (determined by three factors: Leader-member relations, task structure, and the position power) using process evaluation. Findings from the evaluation could be used to update the theory and add contextual nuance. Not only would better theory about leadership in different contexts be useful for other school-based interventions, but it could also help inform interventions (and new hypotheses) for other multi-level interventions in international development.

Opportunities for CEDIL

Our consultation, scoping review and discussions have led us toward an emergent consensus on both promising ways to improve the capacity of evaluations to help us learn more for new settings, but also an appreciation of many of the challenges associated with realising this vision. CEDIL can have a role in testing approaches to evaluation in international development that use these ideas more formally.

One view that was emergent was that there are methodological challenges facing the use of middle-range theory-building as an explicit tool to guide evaluation design. As we have described above, a middle-range theory must strike a balance between local specificity of constructs, which cannot be applied elsewhere, and being too general as to no longer be useful. Following from the context-centred approach and the fidelity-as-function concepts, described above, how something acts within a system is closely related to how we define the concept itself. Addressing this conceptual challenge will

likely require input from a range of stakeholders, and the generation of guidance to make the conceptual process accessible and actionable.

We found that there was no universally shared approach to representing the causal structures in a theory; there are many ways of doing this -- causal pies, DAGs, chain diagrams, system of structural models -- each with different strengths and weaknesses for evaluation practice. Furthermore, a middle-range theory cannot describe everything and remain useful (or definable). There was a consensus that there could be better guidance on how to represent theory, and where to set justifiable limits. As noted above, one of the challenges that faces the use of middle-range theory is finding reliable markers or indicators about where and when a theory is appropriate, equivalent to identifying when the underlying structures that afford the causal pathways hold. Unfortunately, although guidance is possible, and more investment in theory and concept building, it does not seem likely that any set of methods for this can be delineated. Social science is methodologically broad-church: different ideas are invented, developed, refined and tested in a huge variety of different ways. There is no more a fixed set of methods for identifying markers for when a middle-range theory obtains than there is for conceiving of, developing and testing middle-range theories themselves. More work is needed in this area; perhaps by developing and testing theory over a range of evaluations, it would be possible to develop guidance relating to the concept of markers in international development settings.

Standards of evidence for causal claims can limit the extent to which theories can be tested and improved. Not all evidence is equally valid for making causal claims, and our call for more focus on theory and for evaluators to draw on process evaluation does not contradict the generally accepted principles that different designs are better than others for overcoming bias and confounding. It will be essential that the strengths and weaknesses of different sources of evidence are made clear when theories are being tested. Process data, qualitative research, and small n case studies are not a cheap work-arounds for slower and more expensive large-scale designs such as cluster randomised trials. We strongly advocate for the best and most robust design that is appropriate for testing hypotheses implied by theory, while recognizing that this may lead to slower, but surer, advancement in knowledge. Since interventions in different sectors of development may share more than one underpinning theory, it is possible that a slow start will lead to rapid gains in theoretical knowledge.

A challenge is that with complex interventions with multiple components, or any interventions that are expected to act through long causal networks, evaluators may experience an unmanageable number of possible combinations of configurations of each element for any one case. In our illustration of a deworming intervention to improve cognitive outcomes, if each element of the simple causal chain was binary then that still amounts to 2^9 = 512 possible configurations. There are few applications of promising analytical methods, such as mediation, that can handle multiple interrelated pathways (Steen et al., 2017). Theory-led investigation, and qualitative data collection may help to focus evaluation effort to answer questions about the most important variation, for example by using case-studies to explore positive deviants. Focusing on testing hypotheses implied by the middle-range theories behind each of the steps may be able to untangle the many configurations to produce useable learning for elsewhere.

Another challenge for this theory-focused approach is that the incentives for building theoretical knowledge may not be sufficiently strong in some areas of academic research. There have been calls to identify 'what works' in fields from public health, education, gender-based violence, and economic development. The 'what works' agenda may be incentivizing researchers to conduct as many evaluations, usually trials, as possible so that they can maximise their contribution to sorting what 'works' from what does not 'work'. This perspective may resist focusing on theory since the adoption of rigorous experimental methods has been seen as a refreshing departure from the theoretically embedded and grounded research of the past. 'What works' has helped to simplify the policy agenda, democratizing evaluation by relegating substantive knowledge. Relative to results of RCTs, there will be challenges to communicating abstract theoretical concepts to policy makers and convincing skeptical stakeholders that 'what works' (even 'what works for whom') is not asking the questions that allow us to learn more for elsewhere.

Finally, if using middle-range theory is to be taken seriously as a core element of learning more from evaluations, we will need clarity about what makes one theory better -- and more useful -- than another. For example, imagine we find that a particular set of activities gives rise to changes in outcomes in cities but not in rural areas, leading us to theorise that urban/rural is an important underlying factor. This might not be very helpful because are too many differences between urban and rural environments to help make precise predictions for elsewhere and test the theory. It might be more useful to know which of the differences are important (e.g. crime rates, income levels, road transportation, mass transit, building materials, pollution, resources, density, trees), and how. Ever-deepening explanation of the underlying processes is a tension at the heart of the middle-range theory, at once becoming more detailed, but retaining sufficient conceptual flexibility to allow the theories to be applied in other places. The issue of what makes a better theory comes back to the challenge we started with of engaging with a range of disciplines and with decisions makers to ensure that they are useful.

Recommendations

Our review of the literature and consultations have led to recommendations for how CEDIL can strengthen evaluations to 'learn more for next time'. CEDIL should focus on a number of evaluations and systematic reviews as exemplars of evaluation contributing to theory building. These should be in areas that are important in policy terms and have existing theory that is informed by evidence. An evaluation should start with a synthesis of relevant theory. These would be used to inform intervention theories of change or generate hypotheses about how the intervention is expected to function and produce change. The evaluation could then be designed in a way that allows for these hypotheses to be tested, encompassing a range of contexts (place and population), producing results that are useful in strengthening or refining the existing theories that allow for predictions in the contexts of interest. These recommendations are not to be implemented as a stand-alone and generic exercises and should be linked to the preparatory phases of a number of focused empirical studies. We recommend the following specific activities:

Create a Gap Map of theories in development. Gap maps have been used to identify where more research is needed about the effectiveness of interventions, or where few interventions have been evaluated for particular outcomes. A Gap Map for theories in development would compile a taxonomy of the main theories that are used as the foundation of intervention design decisions, and comment on how much is known about each theory. The Gap Map could be used to focus resources on less-well-developed theories in future evaluations. The Gap Map should be developed for practice and viewed as a preparatory step in the design of real prospective evaluations.

Form Theory Hubs. Focusing evaluation on theory rather than on interventions will require stronger interdisciplinary communication and collaboration. CEDIL should actively engage this challenge by looking outside of the development and public health disciplines to build hubs of experts who are deeply engaged with relevant theoretical bodies of knowledge, for example from business schools or marketing firms. To ensure a focus on practice and reality, the themes of the Hubs should be guided by prospective and on-going evaluations.

Implement theory-focused synthesis methods. Theory-based synthesis methods would address questions about theories and draw from across disciplines (e.g. from management to political science) rather than being tied to specific outcomes or interventions (as is the norm). The Gap Map will help identify focus areas, however we expect that theories of implementation, i.e. 'implementation science', will have a lot to offer policy improvement for development, and that there is a lot to learn. CEDIL should identify 2-3 key theory questions about the implementation of contemporary interventions and commission reviews that systematically review the available science to generate questions for planned evaluations.

Apply existing and novel methods to represent theory. While logic models and theories of change are useful communication tools designed to build common ground between various stakeholders, they are not scientific. Scientific theories are either described in narrative argument or using mathematics. Middle range theories draw on multiple disciplines and will change over time. In the context of planning and designing prospective evaluations, CEDIL should support the application and development of methods for the representation of middle-range theory and for putting 'theory' into theories of change.

Pilot theory-based evaluation. We have found that to learn more from evaluations we need to view the primary purpose of the evaluation to be the strengthening of theory. Design decisions from variation of the intervention, focus of the process evaluations, choice of places to collect data, and methods of exploring case studies can all be guided by a new focus on theory and not on interventions. CEDIL should support a suite of studies in substantive areas of important to contemporary international development policy that try out these ideas, as well as a process of reflection and for producing guidance for best practices from this work. The approach should itself be based in epistemological theory and tested through in-depth evaluation in a diverse selection of case-study evaluations.

Identify markers. A key challenge for using theory to guide intervention design is quickly identifying when theories will apply and will be useful for making predictions

about the effects of interventions. Mapping each context, which could be small scale (e.g. families or even individuals) or large scale (e.g. countries) is not feasible. Ideally it would be possible to identify 'markers' in new settings that would inform how the theory should be applied. CEDIL should incorporate this thinking in substantive evaluations and reflect on the challenges and possible solutions in detail.

References

Athey S and Imbens GW (2015) Machine Learning Methods for Estimating Heterogeneous Causal Effects. *stat* 720(101): 1–9. DOI: 10.1073/pnas.1510489113.

Attanasio O, Meghir C and Santiago A (2011) Education choices in Mexico: using a structural model and a randomised experiment to evaluate PROGRESA. *The Review of Economic Studies* 79(1). Oxford University Press: 37–66. DOI: 10.2307/41407044.

Bareinboim E and Pearl J (2013) A General Algorithm for Deciding Transportability of Experimental Results. *Journal of Causal Inference* 1(1). DOI: 10.1515/jci-2012-0004.

Bareinboim E and Pearl J (2014) Transportability from Multiple Environments with Limited Experiments: Completeness Results. *Advances in Neural Information Processing Systems* 27(November): 280–288. Available at: http://papers.nips.cc/paper/5536-transportability-from-multiple-environments-with-limited-experiments-completeness-results (accessed 30 November 2017).

Bates M and Glennerster R (2017) The Generalizability Puzzle. *Stanford Social Innovation Review*. Available at: https://www.povertyactionlab.org/sites/default/files/documents/L8_Generalizability_Bates_2017_0.pdf (accessed 22 June 2018).

Birch S (1997) As a matter of fact: evidence-based decision-making unplugged. *Health Economics* 6(6). Wiley Subscription Services, Inc., A Wiley Company: 547–559. DOI: 10.1002/(SICI)1099-1050(199711)6:6<547::AID-HEC307>3.0.CO;2-P.

Blattman C (2016) Why what works is the wrong question - Evaluating ideas not programs. URL: https://chrisblattman.com/2016/07/19/14411/

Bonell C, Oakley A, Hargreaves J, et al. (2006) Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *BMJ (Clinical research ed.)* 333(7563). British Medical Journal Publishing Group: 346–9. DOI: 10.1136/bmj.333.7563.346.

Bonell C, Fletcher A, Morton M, et al. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science and Medicine* 75(12): 2299–2306. DOI: 10.1016/j.socscimed.2012.08.032.

Bonell C, Mathiot A, Allen E, et al. (2017) Initiating change locally in bullying and aggression through the school environment (INCLUSIVE) trial: Update to cluster randomised controlled trial protocol. *Trials* 18(1). BioMed Central: 238. DOI: 10.1186/s13063-017-1984-6.

Burchett H, Umoquit M and Dobrow M (2011) How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *Journal of Health Services Research and Policy*. DOI: 10.1258/jhsrp.2011.010124.

Burchett HED, Mayhew SH, Lavis JN, et al. (2013) When can research from one setting be useful in another Understanding perceptions of the applicability and transferability of research. *Health Promotion International* 28(3): 418–430. DOI: 10.1093/heapro/das026.

Burchett HED, Mayhew SH, Lavis JN, et al. (2015) The usefulness of different types of health research: Perspectives from a low-income country. *Evidence and Policy* 11(1): 19–33. DOI: 10.1332/174426514X13990430410723.

Cameron S V. and Taber C (2004) Estimation of Educational Borrowing Constraints Using Returns to Schooling. *Journal of Political Economy* 112(1). The University of Chicago Press: 132–182. DOI: 10.1086/379937.

Cartwright N (2017) Single case causes: What is evidence and why. In: *Philosophy of Science in Practice*. Springer, pp. 11–24.

Cartwright N (forthcoming, 2019) Nature the Artful Modeler: Lectures on Laws, Science, How Nature Arranges the World, and How We Can Arrange It Better. Open Court Publishing, Chicago

Castro Felipe González, Barrera, Jr. M and Martinez, Jr. CR (2004) The Cultural Adaptation of Prevention Interventions: Resolving Tensions Between Fidelity and Fit. *Prevention Science* 5(1). Kluwer Academic Publishers-Plenum Publishers: 41–45. DOI: 10.1023/B:PREV.0000013980.12412.cd.

Dane A V and Schneider BH (1998) Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review* 18(1): 23–45. DOI: 10.1016/S0272-7358(97)00043-3.

Davidson EM, Liu JJ, Bhopal R, et al. (2013) Behavior change interventions to improve the health of racial and ethnic minority populations: A tool kit of adaptation approaches. *Milbank Quarterly* 91(4): 811–851. DOI: 10.1111/1468-0009.12034.

De Silva MJ, Breuer E, Lee L, et al. (2014) Theory of Change: A theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials* 15(1). BioMed Central: 267. DOI: 10.1186/1745-6215-15-267.

Deaton A (2009) Instruments of Development: Randomization in the Tropis, and the Search for the Elsuive Keys to Economic Development. *NBER Working Paper Series* (14690). Cambridge, MA: 123–160. DOI: 10.3386/w14690.

Deaton A and Cartwright N (2017) Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*. DOI: 10.1016/j.socscimed.2017.12.005.

Durlak JA and DuPre EP (2008) Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology* 41(3–4): 327–350. DOI: 10.1007/s10464-008-9165-0.

Egan M, Bambra C, Petticrew M, et al. (2009) Reviewing evidence on complex social interventions: Appraising implementation in systematic reviews of the health effects of organisational-level workplace interventions. *Journal of Epidemiology and Community Health* 63(1). BMJ Publishing Group Ltd: 4–11. DOI:

10.1136/jech.2007.071233.

Egger, Matthias, Smith, George Davey, Altman DG (2001) Systematic Reviews in Health Care : Meta-analysis in Context. BMJ Books: 487.

Elster J (2007) *Explaining Social Behaviour. More Nuts and Bolts for the Social Sciences*. Cambridge University Press. DOI: 10.1017/CBO9781107415324.004.

Fairlie RW (2005) An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement* 30(873). North-Holland: 305–316. DOI: http://iospress.metapress.com/content/0747-9662/.

Flowers P, Hart GJ, Williamson LM, et al. (2002) Does bar-based, peer-led sexual health promotion have a community-level effect amongst gay men in Scotland? *International Journal of STD and AIDS* 13(2): 102–108. DOI: 10.1258/0956462021924721.

Funk S, Camacho A, Kucharski AJ, et al. (2016) Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics*. Elsevier. DOI: 10.1016/j.epidem.2016.11.003.

Gaglio B, Shoup JA and Glasgow RE (2013) The RE-AIM framework: A systematic review of use over time. *American Journal of Public Health* 103(6). American Public Health Association: e38-46. DOI: 10.2105/AJPH.2013.301299.

Gardner F, Montgomery P and Knerr W (2016) Transporting Evidence-Based Parenting Programs for Child Problem Behavior (Age 3–10) Between Countries: Systematic Review and Meta-Analysis. *Journal of Clinical Child & Adolescent Psychology* 45(6). Routledge: 749–762. DOI: 10.1080/15374416.2015.1015134.

Garnett GP, Cousens S, Hallett TB, et al. (2011) Mathematical models in the evaluation of health programmes. *The Lancet*. DOI: 10.1016/S0140-6736(10)61505-X.

Gross N (2009) A pragmatist theory of social mechanisms. *American Sociological Review* 74(3). SAGE PublicationsSage CA: Los Angeles, CA: 358–379. DOI: 10.1177/000312240907400302.

Haidt J (2003) The moral emotions. *Handbook of affective sciences* 11(2003). Oxford: Oxford University Press: 852–870.

Hargreaves J, Hatcher A, Busza J, et al. (2011) What happens after a trial? Replicating a cross-sectoral intervention addressing the social determinants of health: the case of the Intervention with Microfinance for AIDS and Gender Equity (IMAGE) in South Africa. *Social determinants approaches to public health. World Health Organization, Geneva*: 147–159.

Hawe P, Shiell A and Riley T (2009) Theorising interventions as events in systems. In: *American Journal of Community Psychology*, June 2009, pp. 267–276. DOI: 10.1007/s10464-009-9229-9.

Hey SP (2014) Theory Testing and Implication in Clinical Trials. Available at: http://philsci-archive.pitt.edu/11045/1/biomarker_theory_testing-psa.pdf (accessed 30 November 2017).

Hill AB (1965) The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 58(5). Royal Society of Medicine Press: 295–300.

Hoffmann TC, Erueti C and Glasziou PP (2013) Poor description of non-pharmacological interventions: Analysis of consecutive sample of randomised trials. *BMJ (Online)* 347(7924). British Medical Journal Publishing Group: f3755. DOI: 10.1136/bmj.f3755.

Hoffmann TC, Glasziou PP, Boutron I, et al. (2014) Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 348. BMJ Publishing Group Ltd. DOI: 10.1136/bmj.g1687.

Hotz VJ, Imbens GW and Mortimer JH (2005) Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* 125(1–2 SPEC. ISS.). North-Holland: 241–270. DOI: 10.1016/j.jeconom.2004.04.009.

Humphreys M, Jacobs A, Aronow P, et al. (2015) Mixing Methods: A Bayesian Approach. *American Political Science Revie.* 109(4). DOI: 10.1017/S0003055415000453

Kelly JA, Murphy DA, Sikkema KJ, et al. (1997) Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. *Lancet* 350(9090). Elsevier: 1500–1505. DOI: 10.1016/S0140-6736(97)07439-4.

Leijten P, Melendez-Torres GJ, Knerr W, et al. (2016) Transported Versus Homegrown Parenting Interventions for Reducing Disruptive Child Behavior: A Multilevel Meta-Regression Study. *Journal of the American Academy of Child and Adolescent Psychiatry* 55(7): 610–617. DOI: 10.1016/j.jaac.2016.05.003.

Leroy JL, Ruel M and Verhofstadt E (2009) The impact of conditional cash transfer programmes on child nutrition: a review of evidence using a programme theory framework. *Journal of Development Effectiveness* 1(2). Taylor & Francis: 103–129. DOI: 10.1080/19439340902924043.

Leviton LC (2017) Generalizing about Public Health Interventions: A Mixed-Methods Approach to External Validity. *Annual Review of Public Health* 38(1): 371–391. DOI: 10.1146/annurev-publhealth-031816-044509.

Lipsey MW (2009) The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders*. Taylor & Francis Group. DOI: 10.1080/15564880802612573.

Lipton P (2004) *Inference to the best explanation*. Routledge/Taylor and Francis Group.

Manguvo A and Mafuvadze B (2015) The impact of traditional and religious practices on the spread of Ebola in West Africa: time for a strategic shift. *The Pan African medical journal* 22 Suppl 1(Suppl 1). African Field Epidemiology Network: 9. DOI: 10.11694/pamj.supp.2015.22.1.6190.

Merton RK (1967) *Social theory and social structure*.

Mihalic S (2002) The Important of Implementation Fidelity. *Journal of Chemical Information and Modeling*: 1–16. DOI: 10.1017/CBO9781107415324.004.

Moore G, Audrey S, Barker M, et al. (2008) Process evaluation of complex interventions: UK Medical Research Council guidance.: 1–134. DOI: 10.1136/bmj.h1258.

Nevin JA and Shahan TA (2011) Behavioral momentum theory: equations and applications. *Journal of applied behavior analysis* 44(4). Society for the Experimental Analysis of Behavior: 877–95. DOI: 10.1901/jaba.2011.44-877.

Oliver R and Skinner CH (2002) Applying Behavioral Momentum Theory to Increase Compliance. *Journal of Applied School Psychology* 19(1). Taylor & Francis Group : 75–94. DOI: 10.1300/J008v19n01_06.

Pawson R and Tilley N (1997) An Introduction to Scientific Realist Evaluation. In: *Evaluation for the 21st Century: A Handbook*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., pp. 405–418. DOI: 10.4135/9781483348896.n29.

Pawson R, Greenhalgh T, Harvey G, et al. (2005) Realist review - A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research and Policy*. SAGE PublicationsSage UK: London, England. DOI: 10.1258/1355819054308530.

Pearl J and Bareinboim E (2014) External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science* 29(4): 579–595. DOI: 10.1214/14-STS486.

Powell J (2011) *The Lessons of Northern Ireland*. London. Available at: http://www.lse.ac.uk/ideas/research/reports/lessons-northern-ireland.

Pronyk PM, Hargreaves JR, Kim JC, et al. (2006) Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: a cluster randomised trial. *Lancet* 368(9551). Elsevier: 1973–1983. DOI: 10.1016/S0140-6736(06)69744-4.

Rivers C, Lofgren E, Marathe M, et al. (2014) Modeling the Impact of Interventions on an Epidemic of Ebola in Sierra Leone and Liberia. *PLoS Currents*. Public Library of Science. DOI: 10.1371/currents.outbreaks.fd38dd85078565450b0be3fcd78f5ccf.

Schorfheide F and Wolpin KI (2016) To hold out or not to hold out. *Research in Economics* 70(2). Cambridge, MA: 332–345. DOI: 10.1016/j.rie.2016.02.001.

Schwartz D and Lellouch J (1967) Explanatory and pragmatic attitudes in therapeutical trials. *Journal of chronic diseases* 20(8): 637–48.

Scott B, Curtis V, Rabie T, et al. (2007) Health in our hands, but not in our heads: understanding hygiene motivation in Ghana. *Health Policy and Planning* 22(4). Oxford University Press: 225–233. DOI: 10.1093/heapol/czm016.

Shadish WR, Cook TD and Campbell D (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Available at: https://pdfs.semanticscholar.org/9453/f229a8f51f6a95232e42acfae9b3ae5345df.pdf (accessed 30 November 2017).

Sparkman G and Walton GM (2017) Dynamic Norms Promote Sustainable Behavior, Even if It Is Counternormative. *Psychological Science* 28(11). SAGE PublicationsSage CA: Los Angeles, CA: 1663–1674. DOI: 10.1177/0956797617719950.

Steen J, Loeys T, Moerkerke B, et al. (2017) Flexible Mediation Analysis With Multiple Mediators. *American Journal of Epidemiology* 186(2): 184–193. DOI: 10.1093/aje/kwx051.

Strevens M (2012) Ceteris paribus hedges: Causal voodoo that works. *Journal of Philosophy* 109(April): 1–35. DOI: 10.5840/jphil20121091138.

Susukida R, Crum RM, Ebnesajjad C, et al. (2017) Generalizability of findings from randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction* 112(7): 1210–1219. DOI: 10.1111/add.13789.

Thompson SG and Higgins JPT (2002) How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 21(11). John Wiley & Sons, Ltd.: 1559–1573. DOI: 10.1002/sim.1187.

TRANSFER framework: Norwegian Institute of Public Health (2017). URL: https://www.fhi.no/en/projects/transfer-rammeverk-prosjektbeskrivelse/

Wang S, Moss JR and Hiller JE (2006) Applicability and transferability of interventions in evidence-based public health. *Health Promotion International* 21(1): 76–83. DOI: 10.1093/heapro/dai025.

Whitbeck C and Bhaskar R (1977) A Realist Theory of Science. *The Philosophical Review* 86(1). Verso: 114. DOI: 10.2307/2184170.

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

White RG, Orroth KK, Korenromp EL, et al. (2004) Can population differences explain the contrasting results of the Mwanza, Rakai, and Masaka HIV/sexually transmitted disease intervention trials?: A modeling study. *Journal of acquired immune deficiency syndromes (1999)* 37(4). Journal of acquired immune deficiency syndromes (1999): 1500–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15602129 (accessed 2 January 2018).

Wingood GM and DiClemente RJ (2008) The ADAPT-ITT Model: A Novel Method of Adapting Evidence-Based HIV Interventions. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 47(Supplement 1): S40–S46. DOI: 10.1097/QAI.0b013e3181605df1.

Wolpin KI (2013) *The Limits of Inference without Theory*. MIT Press.

Woolcock M (2013) Using case studies to explore the external validity of 'complex' development interventions. *Evaluation* 19(3). SAGE PublicationsSage UK: London, England: 229–248. DOI: 10.1177/1356389013495210.

Young R (2015) Source Similarity and Social Media Health Messages: Extending Construal Level Theory to Message Sources. *Cyberpsychology, Behavior, and Social Networking* 18(9): 547–551. DOI: 10.1089/cyber.2015.0050.

Appendices

Appendix 1: Extraction tool

Title

Authors

Year published

Journal

Disciplinary area

Definition/terms used to describe process of using evidence to make policy

Summary of the particular problem described (what/where is the problem)

Are new methods proposed? If yes, continue

Is the paper methodological, applied (to an example), or both?

Summary of method and which problem it addresses

Role of theory

Strategy for addressing context / complexity

Applicability to wide-range of fields

Feasibility of application in CEDIL programme of work

Other comments

Appendix 2: Stage 1 papers

Title	Authors	Year	Journal	Discipline
A General Algorithm for Deciding Transportability of Experimental Results	Elias Bareinboim and Judea Pearl	2013	Journal of Causal Inference	Causal inference
As a matter of fact: evidence-based decision-making unplugged	Stephen Birch	1997	Health Economics	Health economics
Why "what works?" is the wrong question: Evaluating ideas not programs	Chris Blattman	2016	URL:chrisblattman.com	Economics
Assessment of generalisability in trials of health interventions: suggested framework and systematic review	C Bonell, A Oakley, J Hargreaves, V Strange, R Rees	2006	BMJ	Public Health
The usefulness of different types of health research: perspectives from a low-income country	Helen ED Burchett, Susannah H Mayhew, John N Lavis, Mark J Dobrow	2014	Evidence and Policy	Public Health
A tool to analyze the transferability of health promotion interventions	Linda Cambon, Laetitia Minary, Valery Ridde and François Alla	2013	BMC Public Health	Public Health
Single Case Causes: What is Evidence and Why	Nancy Cartwright	2015	CHESS Working paper	Philosophy of science
Why Mixed Methods Are Necessary for Evaluating Any Policy	Nancy Cartwright	Forthcoming	NA	Philosophy of science
The Cultural Adaptation of Prevention Interventions: Resolving Tensions Between Fidelity and Fit	Felipe Gonź alez Castro, Manuel Barrera, Jr., and Charles R. Martinez, Jr	2004	Prevention Science	Public Health
Behavior Change Interventions to Improve the Health of Racial and Ethnic Minority Populations: A Tool Kit of Adaptation Approaches	Emma M. Davidson, Jing Jing Liu,1 Raj Bhopal, Martin White, Mark R.D. Johnson, Gina Netto, Cecile Wabnitz, and Aziz Sheikh	2013	The Milbank Quarterly	Public Health
Mathematical models in the evaluation of health programmes	Geoffrey P Garnett, Simon Cousens, Timothy B	2011	The Lancet	Public Health

	Hallett, Richard Steketee, Neff Walker			
Innovation configurations: analysing the adaptations of innovations	Halle, Gene F.; Loucks, Susan F.	1978	Procedures for Adopting Educational Innovations Program Research and Development Center for Teacher Education The University of Texas at Austin	Education
Making Evidence from Research More Relevant, Useful, and Actionable in Policy, Program Planning, and Practice	Lawrence W. Green, Russell E. Glasgow, David Atkins, Kurt Stange	2009	American Journal of Preventive Medicine	Public Health
Enhancing the portability of public health intervention review evidence for localised decision-making	Dylan Kneale, James Thomas, Alison O'Mara-Eves, Richard Wiggins	Forthcoming	NA	Public Health
Evolution of the Mid Range Theory of Comfort for Outcomes Research	Katharine Kolcaba	2001	Nursing Outlook	Nursing
Generalizing about Public Health Interventions: A Mixed-Methods Approach to External Validity	Laura C. Leviton	2017	Annual Review of Public Health	Public Health
Evidence Required for Adoption of New Vector Control Methods in Public Health	Jo Lines	2013	Biological and Environmental Control of Disease Vectors	Public Health - Vector Control
Declines in efficacy of anti-bullying programs among older adolescents: Theory and a three-level meta-analysis	David Scott Yeager, Carlton J. Fong, Hae Yeon Lee, Dorothy L. Espelage	2015	Journal of Applied Developmental Psychology	Psychology - bullying
The Primary Factors that Characterise Effective Interventions with Juvenile Offenders: A Meta-Analytic Overview	Mark Lipsey	2009	Victims and Offenders	Criminology
Using Case Studies to Explore the External Validity of 'Complex' Development Interventions	Michael Woolcock	2013	Harvard Kennedy School: Faculty Research Working Paper Series	International Development
The ADAPT-ITT Model A Novel Method of Adapting Evidence-Based HIV Interventions	Gina M. Wingood and Ralph J. DiClemente	2008	J Acquir Immune Defic Syndr	Public Health
Applicability and transferability of interventions	Shuhong Wang,	2005	Health	Public Health

in evidence-based public health	John R. Moss, Janet E. Hiller		Promotion International	
Incorporating Demand And Supply Constraints Into Economic Evaluations In Low-Income And Middle-Income Countries	Anna Vassall, Lindsay Mangham-Jefferies, Gabriela B. Gomez, Catherine Pitt And Nicola Foster	2016	Health Economics	Health Economics
The use of propensity scores to assess the generalizability of results from randomized trials	Elizabeth A. Stuart, Stephen R. Cole, and Catherine P. Bradshaw and Philip J. Leaf	2011	Journal of the Royal Statistical Society	Statistics
Experimental and quasi-experimental designs for generalised causal inference	Shadish, Cook, Campbell	2002	NA	Evaluation
A Schema for Evaluating Evidence on Public Health Interventions	Lucie Rychetnik and Michael Frommer	2002	National Public Health Partnership	Public Health
What can we learn from international comparisons of health systems and health system reform?	Barbara McPake, Anne Mills	2000	Bulletin of the World Health Organization	Public Health - Health Systems
Effectiveness, Transportability, and Dissemination of Interventions: What Matters When?	Sonja K. Schoenwald and Kimberly Hoagwood	2001	Psychiatric Services	Psychology
What is this thing called context (and why does it matter for evaluation)?	Mark Petticrew, Lawrence Moore	Forthcoming	NA	Public Health
Nothing as Practical as a Good Theory	Ray Pawson	2002	Evaluation	Evaluation
Transdisciplinary working to shape systematic reviews and interpret the findings: commentary	Sandy Oliver , Paul Garner, Pete Heywood, Janet Jull, Kelly Dickson, Mukdarut Bangpan, Lynn Ang, Morel Fourman and Ruth Garside	2017	Environmental Evidence	Environmental Science
Predicting the efficacy of future training programs using past experiences at other locations	V. Joseph Hotz, Guido W. Imbens, Julie H. Mortimer	2005	Journal of Econometrics	Economics
Improving child safety: deliberation, judgement and empirical research	Eileen Munro Nancy Cartwright Jeremy Hardie Eleonora Montuschi	2016	Centre for Humanities Engaging Science and Society	Philosophy

Contact us

Centre of Excellence for Development Impact and Learning
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org
@CEDIL2018