To Boldly Go Where No Evaluation Has Gone Before: The CEDIL Evaluation Agenda

Edoardo Masset[1], Howard White[2*]

July 2019

[1] Centre of Excellence for Development Impact and Learning, UK
[2] Campbell Collaboraton, Canada and India
* Centre of Excellence for Development Impact and Learning (CEDIL), London International Development Centre. The authors thank Marcella Vigneri for comments on an earlier version of this paper.

Suggested Citation: Masset E, White H, 2109 To Boldly Go Where No Evaluation Has Gone Before: The CEDIL Evaluation Agenda. CEDIL Paper: London

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

Corresponding Author: Edoardo Masset, email: Edoardo.masset@lidc.ac.uk

Table of Contents

Abstract

This paper introduces the rationale behind the establishment of the Centre of Excellence in Development, Impact and Learning (CEDIL) and the challenges it will address. CEDIL was established with funding from the UK Government through the Department for International Development (DFID), with the aim of developing and supporting new methods and approaches in the evaluation of international development interventions in neglected geographic, and thematic areas. CEDIL has concluded its inception phase, which explored new evaluation methods and approaches with a series of papers, workshops, and consultations. This paper sets out CEDIL's research agenda for the next three years. CEDIL will promote studies in three key areas: 1) unpacking complex interventions with mixed methods approaches, and exploration of mechanisms, 2) generalising evidence through the formulation of mid-level theories of neglected interventions, and 3) promoting evidence use through stakeholder engagement and demonstrating what works within such communication.

Section 1

About CEDIL

The Centre of Excellence for Development, Impact and Learning (CEDIL) was established in 2017 under a five-year grant from the UK Department for International Development (DFID) with the goal of filling major gaps in: the evidence base, evaluation methods, synthesis of evidence, and knowledge translation. More specifically, CEDIL, was set up to achieve the following goals: developing new evaluation methods in neglected thematic areas, commissioning impact evaluations and related research, and promoting evidence use.

CEDIL is composed of both a Research Directorate, which provides strategic direction and technical guidance, and a Programme Directorate, which manages the project. The Research Directorate was formed by a consortium of five leading institutions: the International Initiative for Impact Evaluation (3ie), the Campbell Collaboration, the Centre for Evaluation of the London School of Hygiene and Tropical Medicine, the EDePO at University College London, and the EPPI-Centre. The Research Directorate is led by a core Intellectual Leadership Team (ILT) of 11 internationally recognised experts in impact evaluation. The ILT is supported by seven additional expert members of the Extended Leadership Team (ELT),[1] and the activities are overseen by an Advisory Board.[2] A small Research Directorate is based at the London International Development Centre consisting of a Research Director, a Deputy Director, a Research Fellow, and a Project Coordinator.[3] The Programme Directorate is led by Oxford Policy Management (OPM).

The role of the Research Directorate is to identify and support research programmes, addressing key issues in the evaluation of international development projects using innovative methods. The CEDIL Programme Directorate will issue Requests for Proposals (RfPs) to commission impact evaluation studies, systematic reviews, and academic papers under these programmes of work.

Section 2

Introduction

Since the early 2000s, there has been rapid growth in impact evaluations; defined as studies which assess the differences an intervention makes compared to a no intervention counterfactual.[4] The 3ie database shows that the number of impact evaluations of development interventions grew from around 50 per year in the early 2000s, to over 500 by

[1] Members of the ILT are Orazio Attanasio, Nancy Cartwright, James Copestake, Rick Davies, Ian Goldman, David Gough, James Hargreaves, Manny Jimenez, Audrey Prost, Andrew Pulling, and Chris Roche. The ELT includes: Chris Bonell, Mike Clarke, Macartan Humphreys, Jyotsna Jalan, Shahidur Khandker, Sonya Krutikova, and Sandi Oliver.

[2] Members of the Advisory Board are: David Ameyaw (chair), Annie Duflo, Charlotte Watts, Imran Matin, Ken Chomitz, Kingsley Arkorful, Patricia Rogers, Robert Osei, and Ruth Stewart

[3] Current members of the Research Secretariat are: Howard White (director), Susie Burdett, Edoardo Masset, and Marcella Vigneri.

[4] Most impact evaluation methods have an explicit counterfactual usually with outcomes measured by a control group, though there are exceptions (see White, 2009, for further discussion).

2012 (Sabet et al., 2018). There are now nearly 5,000 impact evaluations in the 3ie database.

Despite this growth, there remain gaps in our knowledge surrounding effective interventions. There are four such gaps needing to be addressed:

- *Thematic and geographic evidence gaps*: themes, sectors, or regions/countries for which commonly used methods may be applied, but for which there are few extant studies. Examples include, climate change, conflict, and humanitarian settings, governance, and infrastructure.

- *Methods gaps for primary studies*: evaluation questions, for which there is no consensus in the international development research community, as to the best available approach to answering the question, partly as some methods such as process tracing, qualitative comparative analysis (QCA), integrative mixed method analysis, and predictive analysis, to give more timely information, have been insufficiently tested. Therefore, there are no guidelines, or commonly used texts for these methods. Method gaps are defined by the context of the question being one in which existing methods cannot be applied. Examples of context are, rapid onset emergencies (no baseline and possible need to use real time monitoring information for evaluation purposes), capacity development to a single agency (a small *n* evaluation question) and complex interventions, which may evolve over time with emergent outcomes.

- *Evidence synthesis gaps*: selecting the most appropriate methods for evidence synthesis. There is a tension between the rigour of the synthesis and the time frame and resources available. There are also synthesis approaches under-used in international development, e.g. structural equation modelling approaches to meta-analysis, as well as a need to strengthen qualitative synthesis, and mixed methods synthesis, to draw on a broader range of evidence.

- *Evidence translation gaps*: producing rigorous evidence is not sufficient to ensure its use, nor are the passive dissemination strategies of the past. CEDIL will develop models of stakeholder engagement to increase the relevance, interpretation, and application of study findings for policy and practice. Policy uptake is also a function of evaluation design and timelines.

CEDIL was established by DFID to address these gaps and to promote evaluations of use to assessing and informing DFID policies and programmes. This paper first elaborates on the above-mentioned gaps, then provides an overview of the planned programmes of work to address the gaps.

Section 3

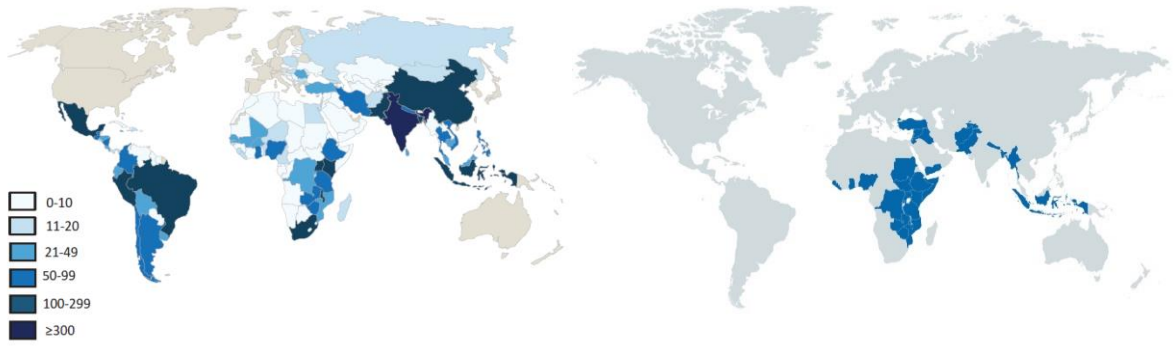
Evidence gaps: CEDIL motivation and priorities

The motivation for establishing CEDIL is the mismatch between the evidence produced by impact evaluations of development interventions today, and the needs of policy makers and practitioners in international development. The mismatch is geographic, thematic, and programmatic. Evidence is lacking for the geographic regions, thematic areas, and project types in which aid agencies are investing most of their resources.

As mentioned above, the last 25 years have witnessed a tremendous increase of impact evaluations in international development. But these impact evaluations are not equally distributed across geographic regions. Nearly half (45%) of all evaluations came from 10 mostly middle-income countries (in order from top to bottom): India, China, Mexico, Kenya, South Africa, Bangladesh, Brazil, Uganda, Pakistan, and Peru. Looking at a higher-level of aggregation, impact evaluations were evenly distributed between the Sub-Saharan Africa region, Latin America, and South Asia, with very few evaluation studies conducted in the Middle East and North Africa region.

The geographic distribution of DFID development assistance is rather different from the geographic distribution of impact evaluations (see Figure 1). DFID has identified 32 priority countries for bilateral assistance. The selection of the countries is based on several factors: i. extreme poverty – which affects a large share of the population; ii. countries' inability to finance poverty reduction; and, iii. specific risks, such as national security priority, the degree of aid to needs ratio, and the ability to deliver programmes. In 2016, over 50% of total UK development assistance went to just 10 countries: Pakistan, Syria, Ethiopia, Nigeria, Afghanistan, Tanzania, Jordan, South Sudan, Sierra Leone, and Somalia. All of these countries are either themselves fragile countries, or neighbour fragile countries. Indeed, DFID has pledged to spend at least 50% of its bilateral budget in fragile states, with the ambition of expanding work in the Middle East, the Sahel and across the Africa 'Arc of Instability' (including Nigeria, the Chad basin, Mali, Niger, Chad, Sudan and Somalia). Illustrating the geographic gap, only Pakistan is in both the list of the top 10 countries with impact evaluations, and the top 10 recipients of UK development assistance.

Figure 1 Geographic distribution of impact evaluations and UK development assistance



Source: Sabet and Brown (2018) (left), and https://www.gov.uk/guidance/where-we-work (right).

The second mismatch is thematic. Nearly half (47%) of all evaluations reviewed by Sabet and Brown was published in health journals. Most evaluations (65%) were conducted in just three thematic areas: health nutrition and population, education, and social protection. Neglected areas, with fewer than 5% of total evaluations, include (in order from least to most neglected): private sector development, water and sanitation, environment and disaster preparedness, ICT, economic policy, urban development, energy, and transportation.

The UK belongs to the small number of countries that meet the commitment of spending 0.7% of their national income in development assistance.[5] Since 2015, the goals of DFID's development assistance have broadened to include: tackling extreme poverty and helping the most vulnerable, promoting global prosperity, strengthening global peace, security and governance; and strengthening resilience and response to crises.[6] Development funding has shifted accordingly, notably in favour of humanitarian assistance, which is now the sector with the largest share of aid spending. The other leading thematic areas, in addition to humanitarian assistance, are health and sanitation, government and civil society, education, and economic infrastructure. An important shift towards interventions promoting improvements in the business environment has also occurred.

The mismatch between evidence from impact evaluations and the practice of development is not limited to thematic and geographic areas of intervention. It is also believed that most impact evaluations today are assessing relatively 'simple' interventions, consisting of simple programme theories, fewer stakeholders, clear goals, and operating in uncomplicated environments. Impact evaluations will typically assess just one component of an intervention, ignoring, for example, its capacity building and institutional development objectives and components. However, most development programmes operate in uncertain and complex environments, and consist of packages of different initiatives promoted by different stakeholders; often without a clear definition of goals or the means to achieving them. This mismatch points to the second gap to be addressed by CEDIL; the methods gap.

A key assumption behind the establishment of CEDIL is that the absence of impact evaluations in difficult geographic, thematic, and programmatic contexts is not the result of a lack of interest on the part of researchers and funders. Rather, it is the consequence of the lack of appropriate approaches and methodologies to conduct 'difficult' impact evaluations. Standard impact evaluation techniques, such as randomised control trials and quasi-experimental designs may not easily be applied to complex interventions, and so are unable to answer all relevant policy questions. There is a need for researchers to devise and employ new methods and approaches, which are fit for purpose and suited to difficult evaluation environments.

Other considerations also constrain the perceived usefulness of impact evaluations, such as questionable external validity and consequent lack of transferability of study findings to other settings, too great a focus on what to do and not how to do it (also phrased as what works not how it works), and untimely information with studies only yielding findings for action after key decision points have passed.

A similar set of challenges faces evidence synthesis. Whilst methods are well established to synthesise effect sizes from 'large n' quantitative impact evaluations, there is less experience in how to synthesise impact estimates from other approaches such as QCA and structural modelling. Similarly, traditional reviews take longer than a commissioner would like; either because of their approach or lack of primary studies. This means that they are

[5] This and the following paragraph are largely based on Baker et al. (2018)

[6] https://www.gov.uk/government/organisations/department-for-international-development/about

unable to report on, to the desired extent, design and implementation issues, or how variations in context may condition impact.

Section 4

Challenges which CEDIL will address

CEDIL will address three fundamental evaluation challenges:

1. The evaluation of complex interventions: development interventions are mostly complex, being composed of interacting parts linked in various ways. This is often ignored, and such interventions are usually evaluated one component at a time or in a black-box fashion, thus ignoring the interactions between components.

2. The transferability of evaluation results: methods to extrapolate the results of evaluations from one context to another are currently underdeveloped. There is great uncertainty with regards to the extent the result of a study can be applied or can be adapted to a different context.

3. The policy impact of evaluations: though the goal of evaluation is informing policy, relatively little is known about the use of evaluation results and how these can be promoted in policymaking.

The first two challenges – complexity of the interventions and transferability of the results – are related. It can be argued that the higher the complexity of the intervention, the lower the transferability of the evaluation's results. The workings of complex interventions are very sensitive to variations in implementation and to the characteristics of the context, consequently the evaluation's external validity of these interventions is very limited. However, solutions to these challenges do not need to be the same. For example, a 'package' of interventions could be successfully evaluated without producing results that are universally valid. Therefore, in our programmes of work we will support studies that are addressing the two challenges separately, though we are aware that in some cases addressing complexity may contribute to our understanding of the transferability of the results, and vice-versa.

This work will take place within the context of CEDIL's list of priority thematic areas, geographic areas, and programme types, which have been identified by DFID in Table 1 below. In its inception phase CEDIL explored the most pressing questions and potential answers related to these difficult contexts, which will be discussed in the next section.

Table 1 DFID thematic, geographic and project priorities

Thematic areas	Geographic areas	Project types
Economic development, inclusive growth, trade, investment, infrastructure and building markets, governance, conflict, extremism, security and stability, justice, anti-corruption including illicit financial flows, humanitarian assistance, migration and modern slavery, climate change and sustainable resource management, accelerating use of family planning, education, disability and gender, nutrition, southern innovation and use of technology	Afghanistan, Bangladesh, Burma, Chad, DRC, Ethiopia, Iraq, Jordan, Kenya, Lebanon, Malawi, Mali, Mozambique, Nepal, Niger, Nigeria, OTP, Pakistan, Rwanda, Sierra Leone, Somalia, South Sudan, Sudan, Syria, Tanzania, Uganda, Yemen, Zimbabwe	Complex interventions, packages of interventions, programme delivery instruments

Section 5

Unpacking complex interventions

In this section we will discuss: 1) What complex interventions are; 2) Why evaluating complex interventions is important; and, 3) How to evaluate complex interventions.

5.1 What are complex interventions?

Complex interventions are commonly understood as interventions consisting of many interacting components (Craig et al., 2006). This definition is very broad and needs to be qualified. Firstly, we introduce a distinction between the *complexity of interventions* and the *complexity of the system*. CEDIL is concerned with the former but not necessarily with the latter. CEDIL has the goal of evaluating *complex interventions,* rather than evaluating interventions implemented in *complex systems*. Some researchers argue that since economic, social, and political systems are complex, their behaviour is uncertain and unpredictable. It is believed that in these systems causes of effects are not just unknown, but cannot be known (CECAN, 2018, Ramalingam et al., 2008). According to this view, standard evaluation methods addressing causality, such as, randomised control trials, are of little use. The Centre for the Evaluation of Complexity Across the Nexus (CECAN) proposes to develop and promote policy evaluations, which are 'complexity-appropriate', employing methods such as agent-based-modelling, qualitative comparative analysis (QCA), system mapping, and Bayesian updating (CECAN, 2018).

In a similar vein, the proponents of the Doing Development Differently movement (ODI, 2016) and others coming from an adaptive management tradition (Wild et al.,2017) have argued that the complexity of social environments is such that results of impact evaluation can rarely, if ever, be generalised. Since the characteristics of contexts and of project implementation vary enormously, standard evaluation methods identifying impacts in a particular context are of little guidance toward policy. This has led some authors to suggest that interventions in complex environments should not be evaluated with traditional methods

(Pritchett et al., 2013) and have proposed replacing mainstream impact evaluation with structural experiential learning, whereby managers adapt interventions to the complexity of the environment through a process of learning-by-doing.

We do not dispute the observation that social systems are complex, and there is much to commend in the concept of using and promoting a plurality of non-traditional methods when evaluating interventions in complex environments. However, CEDIL's primary interest is evaluating *complex interventions,* whether they are implemented in complex systems or not. Some examples below will provide an illustration of what CEDIL means by complex systems and complex interventions.

There is no single accepted definition of a *complex system*, but some characteristics and patterns are typically associated with this notion (Durlauf, 2005): path dependence (when specific events and shocks affect the long-run condition of the system), phase transition (when systems undergo a radical qualitative change for a small change in conditions), and emergent properties (when particular outcomes occur at some higher level of aggregation, for example economies of scale and coordination games). Similarly, CECAN mentions the following characteristics as typical components of complex systems (CECAN, 2018): 'adaptation to changes, feed-back loops, multiple scales, thresholds of change, past states influencing possible future states, being highly dynamic, and being an open system, impossible to bound.' The emergent characteristics of these systems make their behaviour difficult to predict.

As in the case of complex systems, there is no single accepted definition of complex interventions. For example, Funnel and Rogers (2011) distinguish 'simple', 'complicated' and 'complex' interventions, where complicated interventions are characterised by multiple objectives and levels of implementation, whilst complex interventions display emergent properties. The new MRC guidelines on developing and evaluating complex interventions (2019) also state that complexity is not just a characteristic of the intervention, but also of the system in which it is implemented. CEDIL's definition of complex interventions is broad, and includes all interventions consisting of multiple components and implemented as 'packages', whether they display emergent properties or not. Some interventions are characterised by long and complex causal chains. Small variations in the elements of the intervention change the final outcomes, but without displaying emergent properties. For example, growth-monitoring nutrition interventions are notoriously complicated in this sense. The improvement in the nutritional status of a child is the result of a series of carefully timed and sequentially connected activities, each being dependent on several behavioural responses. Small changes in the result chain of the project alter intermediated and final impacts.

Other interventions are complex, in that they produce emergent outcomes, and in this way, the interactions between the project components produce effects that are different from the sum of the effects of each component when implemented separately. These interventions include several components, with the opinion that the solution of the overarching problem requires concerted, and mutually reinforcing investments in different areas whilst they generate a cumulative effect. An example of this type of project is the BRAC ultra-poverty graduation programme, which consists of concerted investments in health, asset transfers, and social protection with the goal of producing a 'big push' in breaking the poverty trap (BRAC, 2018). Integrated rural development projects, which were common in the 1970s, are another example of this type of intervention (Masset, 2018).

Finally, some interventions consist of multiple activities but do not have a unified programme theory. It is noted that a programme theory underlies all interventions, and that projects are 'theories incarnate' (Pawson and Tilley, 2004). But this is not always true, as various projects are the result of a political compromise on the use of resources between the funder, the implementer, and the beneficiaries, in such a way that they comprise of many components,

which are not necessarily linked to each other in a logical way. Home-grown school feeding programmes, whereby two distinct goals (feeding children and supporting local farmers) are brought together, are an example of this type of programme. The two goals are merged in the interest of providing benefits to different stakeholders at the same time, but there is no explicit underlying logic linking the provision of food to school children and sourcing the same food locally. Multi-country interventions, and development portfolios within a country also belong to this category of interventions.

5.2 Why evaluate complex interventions?

Three common approaches to evaluating complex packages of interventions are: 1) black box evaluations; 2) evaluations of single components of the intervention package; and, 3) no evaluation. In the first case, they produce results that have limited applicability outside the area of intervention, whilst in the second case they produce results that can mislead policy. In the third approach, researchers abandon the possibility of conducting an impact evaluation altogether.

Black box evaluations are the default approach to the evaluation of complex interventions. These evaluations assess the impacts of the intervention 'as a package' without investigating the mechanism producing the outcomes and the interactions between project components. Examples of this type of evaluation are the evaluations of the BRAC poverty graduation programme (Banerjee et al., 2015) or the evaluation of the Millennium Village Project (Mitchell et al., 2018). These projects feature multiple components that are designed to interact to produce synergistic effects. For example, in the BRAC poverty graduation programme, the beneficiary receives an asset transfer and a cash transfer to ease adverse conditions whilst they work their way out of poverty. In the Millennium Village Projects, farmers are trained and given free fertiliser to improve their crops, but they also have access to new health services that improve their productivity. Black box evaluations do not explore synergistic effects and how they operate. Evaluations of these programmes can therefore only recommend whether to continue the intervention 'as is' or discontinue it entirely. They do not inform how the programme could, or should be improved, or how to adapt it to different circumstances.

Black box interventions are also poorly informative in the case of complicated projects with long causal chains. As the implementation of a project becomes more complex, this increases the number of choices implementers must make, and the variety of resulting possible outcomes in different contexts. Even an apparently simple project, like a cash transfer programme, requires making several choices; should transfers be given to mothers or fathers? Should they be in cash or kind? Should they be made weekly, monthly, or yearly? It is easy to think how these choices could lead to varying outcomes in different contexts. One main problem of black box evaluations of complicated interventions is therefore the lack of external validity; an issue to which we will come back in the next section when discussing the second CEDIL programme of work. Clearly, the more complex the intervention, the less likely its transferability to different contexts.

It is also common to evaluate one component at a time in complex interventions, or to assume that the intervention consists of just one component whilst ignoring all others. In some cases, an attempt is made to add together evaluations of separate components to assess the impact of the whole intervention. This approach, however, can be misleading. To illustrate this, we borrow some simple notation from Ravallion's discussion of the evaluation of 'portfolio interventions' (Ravallion, 2015). For simplicity, imagine a programme consisting of just two components (x_1 and x_2) affecting some target outcome (y). The programme could be a conditional cash transfer intervention (x_1) implemented together with a nutrition education component (x_2) as in the well-known PROGRESA Mexican cash transfer

programme. The intended outcome is increasing children school attendance. If we assume that the two components affect the outcome directly and independently, the impact of the intervention can be defined:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (1)$$

Evaluating one component at a time can be misleading even in the absence of interactions between components. Nutrition education can have a direct impact on school attendance to the extent that healthier children are less likely to miss school days because of illness. Attributing the impact of PROGRESA just to the cash transfer is not accurate and can overstate the impact of the transfer.

Evaluating one component at a time is even more misleading in the presence of interactions. In the PROGRESA example, nutrition education may encourage the use of more nutritious and more expensive food, so that a cash transfer may enable households to achieve a healthier and more expensive diet. The impact of the intervention with interactions is:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \quad (2)$$

in which the last term on the right is the impact of the two components being implemented together. Ignoring this interaction omits a central feature of how the programme works, leading to wrong inferences on the impact of each component.

Issues related to the evaluation of complex interventions are too many to be discussed here, but we mention the following as examples of common challenges: identifying control groups when projects are implemented at a national level or across several countries; understanding how the programme works when activities are spread across many sectors, and carried out by different implementers; assessing impact when outcomes are many and there is no obvious way to consider them together; and, evaluating the cost-effectiveness of the intervention when outcomes are many.

5.3 How to evaluate complex interventions?

Several approaches have been suggested to evaluate complex interventions, though none of them is fully convincing or applicable to most evaluations. Factorial designs can address synergistic effects, but rapidly run out of statistical power for all but a small number of intervention components and their interactions. The Medical Resource Council guidelines for the evaluation of complex interventions (Craig et al., 2006) recommend experiments and quasi-experiments alongside rigorous process evaluations, which explain how an intervention works, unintended consequences, and causal mechanisms. A mixed-methods (qualitative plus quantitative) approach to evaluation (Jimenez et al., 2018) seems to be a minimum requirement to understand the operation of a complex intervention.

CECAN (2018) recommends the following methods: qualitative comparative analysis, agent-based modelling, dependency models, and Bayesian updating and dynamic pattern synthesis. As seen above, certain methods were originally designed to analyse complex systems, rather than evaluating complex interventions, but they could be useful in some cases. In economics, complexity has long been researched both theoretically and empirically with a variety of methods: structural equation modelling, instrumental variables, and general equilibrium modelling.

CEDIL is not prescribing which of these methods are to be used, but rather inviting innovative approaches to evaluate the impact of complex interventions.

Section 6

Generalising and transferring evidence through mid-level theories

In this section we discuss: 1) The fundamental problem of generalising and transferring evidence from evaluations; 2) What are mid-level theories? And, 3) How can mid-level theories inform impact evaluations?

6.1 The problem of generalising and transferring evidence

One of the goals of impact evaluations is to produce evidence to inform decisions in settings that are different from the one in which the intervention was originally implemented. For example, if an impact evaluation has found a significant impact of a school feeding intervention in a remote area of a poor country, will the same project have the same or a similar effect in another country? At another time? What can we expect if the project is scaled up at the national level? What is the impact if the intervention is modified? There is no simple answer to these questions, and in some cases, there is no answer at all.

The results of an impact evaluation hold for a specific population, for a project with specific characteristics, when implemented in a particular way, and under the environmental conditions prevailing at a point in time. There is no guarantee that the same results will materialise in other settings, when the project is implemented in a different way, or if external conditions have changed. The higher the sensitivity of the results to project and population characteristics (and the larger the number of such characteristics) the less likely the results of one study will be transferable to other contexts. The complexity of an intervention, as previously noted, may imply limited generalisability or transferability.

Generalising can be interpreted in two ways: 1) as the problem of finding some universal knowledge about the impact of an intervention (generalisability), and 2) as the problem of finding knowledge about the impact of the intervention in a particular case (transferability). To illustrate the point one can think of a universal project effect as the average of all possible effects in all contexts. The impact estimated by a single evaluation is then one sample realisation of the universal effect. A series of impact evaluations can show that an intervention works ‘in general’, in the same way a single intervention can show that an intervention ‘works’ for the average beneficiary. From a policy perspective, it is equally important to know whether the impact found in a single study, or in a number of studies, will also occur in another context. In most practical policy applications, we would like to know how a specific intervention would work, not in general, but in a given context. What is the impact evaluation from a single study telling us about the impact of the same project in a different context?

The distinction between generalisability and transferability matters because the latter is far more likely to be achievable. A generalisable finding is one which is universally valid; it holds in all contexts. A transferable finding can be transferred from one setting to another. The transferability depends on which contextual factors matter for transferability and is therefore conditional on the presence of those factors. Contextual factors which do not condition the effect can vary (or not) – it does not matter.

The quality and rigour of a study does not improve alone its general validity. The results of a well conducted randomised control trial are silent on general validity as any other quasi-experimental study (Deaton and Cartwright, 2018). The statistical literature has developed several methods to address this problem. Multisite impact evaluations are one example of this approach. When one experiment is not sufficient to produce universally valid results, perhaps many experiments can be conducted in different contexts, and then the effects are summarised. Multisite experiments conducted in different contexts serve this purpose. For example, the multisite evaluation of the BRAC ultra poor poverty graduation programme, found that it works across a range of different contexts of implementation and populations (Banerjee et al., 2015).

Though this type of study 'proves' whether an intervention works in general, it does not tell us what the likely impact of the intervention is in any specific context. Methods to explore the heterogeneity of impacts, and their extrapolation to different contexts, have been proposed and include meta-regression analyses and systematic reviews of evidence, statistical models exploring heterogeneity directly using multi-level modelling, or predictive models based on big datasets. These approaches come at a considerable cost, because they require a large number of impact evaluations, or sophisticated statistical methods. CEDIL accepts the validity or the applicability of these methods, but it proposes to complement them with the adoption of methods that are based on an exploration of the mid-level theory of a particular intervention.

6.2 What is mid-level theory?

There is no standard definition of mid-level theory, and the term may take different meanings in different disciplines. Some general elements of a mid-level theory, can however, be identified (Pawson, 2000). A mid-level theory is not a methodology and is not an empirical regularity. It can be thought of as a 'consolidation of explanations' expressed in the form of 'tendencies' or 'mechanisms'. An example from sociology is the 'reference group theory' of Merton, who invented the term mid-level theory. Reference group theory postulates that people refer to particular groups of other people to make decisions, and to assess their own situation. Reference group theory does not explain all social behaviour, but can explain, for example, some typical spending behaviour.

The attention to 'mechanisms' highlights the similarity of mid-level theory to realist approaches to evaluation. Indeed, the mechanism-context-outcomes approach, used in realist approaches to evaluation, is an operationalisation of mid-level theory. In realist evaluation (Pawson and Tilley, 2004), the researchers aim to understand the process by which people act upon the resources and services offered (the mechanism), in a set of specific economic, social, and other characteristics (the context), which result in consequences, which are at times intended and at times unintended (the outcomes).

One way of understanding mid-level theory, is as a theoretical model or explanation of events that lies between highly contextualised explanations of specific interventions, such as, for example, in the theory of change of a project, and grand social theorising (Davey et al., 2018). Being in the 'middle', a mid-level theory does not explain everything, like grand social theories, but explains more than a specific theory of change.

One way of conceptualising mid-level theories is therefore as nested models:

$$y = f(x) \text{ nested in } y = f(x, h) \text{ nested in } y = f(x, h, z) \quad (3)$$

The first model on the left-hand side is the most general model, which is valid for any project (*h*) and population (*z*) characteristic. The model on the right-hand side is the most fully specified model; it spells out how population and project characteristics may affect the

success of the intervention, for example in a fully specified theory of change. The model in the middle is a mid-level theory. The mid-level theory is implied by the more general model, but the reverse is not true. Similarly, the mid-level theory implies the fully specified theory of change, but the reverse is not true. A mid-level theory as defined above explains how a program works in a plurality of contexts, and therefore implicitly generalises. For CEDIL, mid-level theory has come to represent a renewed interest in the theory behind the intervention, rather than on its effectiveness.

One example of how mid-level theorising leads to generalisation can be found in Cartwright and Hardie (2012) and draws on the evaluation of the Bangladeshi Integrated Nutrition Project (BINP). BINP was a complex nutrition intervention implemented by the World Bank in Bangladesh in the late 1990s, after the successful implementation of a similar project earlier on in Tamil Nadu. A key component of the intervention was the provision of counselling to mothers at periodic growth-monitoring sessions, in which children's height and weight were measured. Child growth lower than prescribed would trigger the provision of nutrition education advice on matters, such as breastfeeding and nutritious diets. In the evaluation of the intervention, White and Masset (2007) used an equation similar to (4) to assess the impact of the intervention:

$$Z_i = \alpha + \beta P + \sum_{j=1}^{k} \gamma_j X_j + \varepsilon_i \quad (4)$$

in which the dependent variable $Z_i$is the nutritional status of child *i*, *P* is the intervention, i.e. 'mothers' counselling', and the X_s are covariates determining nutrition and unaffected by the intervention, such as mothers' baseline social status and education. The parameter β identifies the impact of the intervention. Whilst the literature suggests that the project was successful in Tamil Nadu ($\beta > 0$), the authors found limited impact of the intervention in Bangladesh ($\beta = 0$).

One possible explanation for the lack of impact in Bangladesh is that the populations in the two countries are vastly different and therefore are likely to respond differently to the same treatment. For example, families in Bangladesh could be much poorer than families in Tamil Nadu, hence unable to purchase the nutritious food recommended in the counselling sessions. In the absence of overlap between the covariates that mediate the impact of the intervention, the extrapolation of effects from one country to the other is not possible.

However, a more fundamental reason for the absence of impact in Bangladesh could be the lack of effectiveness of the postulated mechanism ('mothers' counselling'), regardless of the covariates. Conceivably, in Bangladesh, mothers have no say about what food the family purchases and what the children eat. Indeed, the authors (White and Masset, 2007) found that the programme was particularly unsuccessful in families in which mothers and mothers-in-law were living together, presumably because in these more traditional families it is the mothers-in-law and husbands who make decisions about food expenditure and consumption.

Nonetheless, note that if we were to replace 'mothers' counselling' in *P* with 'counselling the person with decision power in the household', then equation (4) and the project may be found to work in both countries. 'Counselling the person with decision power in the household' is at a higher level of abstraction than 'mothers' counselling'. It is mid-level because it explains how children's diets can be improved without specifying the details of which person in the family should be counselled. The person with decision making power in the family varies from one context to another, and the project should be adapted to reflect this understanding of how the family dynamic operates.

This is only one example of mid-level theorising and it is not meant to capture what a mid-level theory is. In this example, the mid-level theory was developed with the support of

qualitative research in the field on relationships within households, and with exploratory data analysis, but other approaches to mid-level theorising are possible, as discussed below.

6.3 How do mid-level theories inform impact evaluations?

How can a mid-level theory be designed, and can it be tested with the data? As mentioned, mid-level theory is not a research method or an empirical generalisation. The approach to formulate mid-level theories implies theorising and conceptualising how projects work. There is no 'how-to' guide to build mid-level theories.

The goal of formulating a mid-level theory is primarily to ask the right questions. Evaluation questions should be framed at a sufficient level of generality to explain a range of phenomena. The identification of tools and methods to evaluate projects is of secondary concern. Mid-level theories can be developed with a combination of approaches including: reviews of evidence, such as the general theories of change that can be found in some systematic reviews; qualitative research on how projects work; exploratory data-analysis investigating the heterogeneity of impacts across populations and intervention characteristics; and, consultations with stakeholders, including project implementers, beneficiaries, and people involved in the intervention.

Mid-level theories cannot be tested with the data and cannot be proved to be right or wrong. However, once formulated, its theoretical implications can be studied, observed empirically, and tested. Testing the implication of a mid-level theory will not lead to its acceptance or rejection, rather it will allow a refinement of the theory, and a better understanding of how the intervention mechanism works.

The theory behind some types of new interventions, for example, humanitarian assistance and anti-slavery projects, is relatively under-developed, so mid-level theorising in these under-researched thematic areas can be done through reviews of evidence, behavioural modelling, field observation, and stakeholder consultations. CEDIL intends to commission work both to elaborate how to develop and use mid-level theories, and work with applications to develop mid-level theories for a range of interventions of interest to DFID, such as support to elections. CEDIL is also interested in work which applies such approaches to develop and test 'transferability frameworks'.

Section 7

Promoting evidence use

This section discusses what we know about the use of evaluation findings, what CEDIL will do to promote the use of evidence produced by the Centre, and to advance knowledge in the promotion of policy impact.

7.1 What do we know about evaluation use?

Informing policy is the goal of programme evaluation. Evaluations can influence policies in different ways. Two main uses of an evaluation are highlighted in the literature: instrumental and conceptual (see for example, Alkin et al, 2017, Cousins et al, 2002, and Nutley et al, 2008). In the instrumental use, evaluation results are used to make immediate decisions, for example, about discontinuing, modifying, or scaling-up an intervention. In the conceptual

use, results of evaluations change policy-makers' perspectives about a specific issue, or causal mechanisms, sometimes in the long-term and in a subtle way. There is also a 'process' use of evaluation, which occurs whilst an evaluation study is implemented, rather than at the end of, the project. In the process use, an evaluation can improve the skills of project managers, and other stakeholders, and provide useful information whilst the programme is still ongoing (Herbert, 2014). There are also potential misuses of evaluations and 'symbolic' uses: when studies are undertaken to fulfil a requirement, or to validate decisions made (McNulty, 2012).

Researchers have identified several factors associated with evaluation use. For example, Cousins et al. (1986) mention relevance, credibility, evaluation quality, communication quality, finding and timeliness, information needs, decision characteristics, political climate, competing information, personal characteristics, and commitment/receptiveness to evaluation. Johnson et al. (2009) include stakeholders' engagement and evaluators' competence. Conceptual frameworks of evidence use have been framed as 'loose' theories of change (Davies, 2016). Researchers often identify the main factors affecting evidence use, but without spelling out the how they operate, and without providing empirical evidence. Two CEDIL inception papers (Oliver et al., 2018, and Brown et al., 2018) review the factors which affect policy uptake of research findings, and how this can be increased.

Recent years have seen the emergence of knowledge brokering, and as a discrete activity within organizations created for this purpose. This is most evident in the What Works movement, that is the What Works clearing houses in the United States and the What Works Centres in the United Kingdom (see Gough et al., 2018, White, forthcoming and Gough and White, forthcoming) for a discussion. The What Works movement has developed evidence portals.

White (forthcoming) distinguishes between evidence databases, platforms, portals, guidelines, and checklists. The former are collections of studies on a specific theme, such as the 3ie database as a collection of studies on the effectiveness of development interventions, and Epistimonikos on health. Evidence platforms are a collection of evidence resources, such as links to recent papers, blogs and relevant evidence collections including databases. Evidence portals, guidelines and checklists are evidence-based decision-making tools, which enable decision-makers to make decisions without having to consult the original research papers. They vary in the degree of agency given to the decision-maker. Portals, such as the Education Endowment Foundation's Teacher and Learning Toolkit, make the evidence available to inform decisions. Guidelines, of the sort produced by the World Health Organization, make recommendations based on expert review of the evidence, and checklists present a 'just do this' approach to key decisions in programme design or implementation. However, outside of health, these approaches have been little used in international development.

7.2 How CEDIL will promote evidence use

CEDIL identified three promising areas of research on evidence use: stakeholder engagement, making sense of evidence, and developing and testing communication methods.

There is enough evidence to suggest that stakeholder engagement, under the right circumstances, can improve evidence use by policymakers (Oliver et al, 2018). Stakeholders make better decisions when they engage with research findings at an early stage, and early engagement favours research that considers stakeholders' needs and interests. The influence of political factors, democratic processes, institutional mechanisms, values, and priorities is however not well understood. CEDIL will promote the formulation of new models of stakeholder engagement, which a) are based on a good understanding of the multiplicity

of interests, attitudes, beliefs and institutional contexts of different stakeholders, and b) that operate a distinction between engagement with knowledge production and engagement with evidence use, as these obey to very different processes, and which explicitly takes into account the uncertainties in decision-making.

Policymakers are often confronted by multiple sources of evidence. Evidence is rarely produced in a vacuum. There are often different types of evidence available for any specific issue and of different quality, and the perceptions of the quality of different pieces of evidence may also contrast. Various sources of evidence may provide conflicting recommendations on the same issue or can contradict other factors affecting evaluation use. There are no easy recommendations on how policymakers should navigate the available evidence and making sense of conflicting information. CEDIL will work at producing guidelines for policymakers on how to use evidence from multiple sources for informed prediction and decisions. The guidelines will take the form of 'rules of thumb' to make decisions easier, more systematic, and rational.

Finally, CEDIL will promote the development of different communication methods in areas of importance to DFID. These may include evidence portals and the development of evidence-based guidance or checklists which were discussed above. Empirical studies assessing the effectiveness of different communication methods will also be supported under CEDIL. There is a sense that traditional dissemination methods are rather ineffective in influencing policy, and that approaches based on establishing networks and personal contacts (knowledge brokering) may be more effective (Brown et al, 2018). However, very little empirical evidence is available on the effectiveness of communication methods of different types. For example, Masset et al. (2012) conducted an experiment to assess the impact of a policy brief on readers' beliefs and attitudes, but this type of study remains extremely rare. Finally, CEDIL will promote innovative empirical studies to generate evidence on what works in communications.

Section 8

Conclusions and next steps

The growth of impact evaluations has greatly contributed to our knowledge of what works, but important evidence gaps remain with respect to geography, themes, appropriate methods, and policy uptake. DFID supported the creation of CEDIL to fill these gaps.

CEDIL is commissioning a series of studies to develop and test innovative methods of impact evaluation and evidence synthesis. These studies fall under three themes: evaluating complex interventions, mid-level theory, and better evidence use. The studies which are commissioned should be of relevance to DFID. This relevance is ensured by thematic and geographic focus, but this relevance is also ensured by inclusion of several evaluations of DFID programmes.

CEDIL is a learning project. Each year it will host an annual event to highlight work in progress and the completed studies it has funded. Lessons-learned papers will be produced, as well as other products to support learning from, adaptation to, and adoption of CEDIL-supported methods.

The current round of studies is the first phase of CEDIL, but the project intends to build on this body of work to contribute to advancing the practice of rigorous impact evaluations of development interventions.

References

Alkin, M.C., and King, J.A. (2017). Definitions of Evaluation Use and Misuse, Evaluation Influence, and Factors Affecting Use, American Journal of Evaluation. 38(3), 434-50.

Baker, A., Crossman, S., Mitchell, I. Tyskerud, Y. and R. Warwick (2018) How the UK spends its aid budget, in Institute of Fiscal Studies. The IFS Green Budget 2018, London: IFS

Banerjee, A., Duo, E., Goldberg, N., Karlan, D., Osei, R., Pariente, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. Science, 348(6236):772-788

BRAC (2018). BRAC's Ultra-Poor Graduation Programme: An end to extreme poverty in our lifetime. BRAC, www.ultrapoorgraduation.org.

Cartwright, N., and J. Hardie (2012) Evidence-Based Policy: A Practical Guide to Doing it Better. Oxford University Press: London

CECAN (2018), Policy evaluation for a complex world, April 2018, Version 2.0. Online at www.cecan.ac.uk

Cousins, J.B. & Leithwood, K.A. (1986) Current Empirical Research on Evaluation Utilisation. Review of Education Research, 56(3), 331-364.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., and M. Petticrew (2006) Developing and evaluating complex interventions. Medical Research Council: London.

Deaton, A., and N. Cartwright (2018) Understanding and misunderstanding randomised controlled trials. Social Science and Medicine, 210C, 2-21

Durlauf, S.N. (2005) Complexity and Empirical Economics. The Economic Journal, Vol. 115(504), F225-F243

Funnel S.C, and P. Rogers (2011), Purposeful Programme Theory: Effective use of theories of change and logic models. Wiley & Sons, San Francisco

Gough D., C. Maidment, and J. Sharples (2018). UK What Works Centres: Aims, methods and contexts. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London. ISBN: 978-1-911605-03-4.

Gough, D. and H. White (forthcoming) 'Using evidence in social policy: From NICE to What Works' in L. Teixeira and J. Cartwright Using Evidence to End Homelessness. London: Policy Press.

Herbert, J.L (2014) Researching Evaluation Influence: A Review of the Literature, Evaluation Review, 38(5), 388-419

McNulty, J. (2012). Symbolic use of evaluation in the international aid sector: arguments for critical reflection. Evidence and Policy, 8(4), 495-509.

Masset, E. (2018) Integrated development, past and present. IDS Bulletin, Vol. 49(4), 17-31

Mitchell, S., Gelman, A., Ross, R., Chen, J., Bari, S., Huynh, U. K., Harris, M. W., Sachs, S. E., Stuart, E. A., Feller, A., Makela, S., Zaslavsky, A. M., McClellan, L., Ohemeng-Dapaah, A., Namakula, P., Palm, C. A., and Sachs, J. D. (2018). The millennium villages project: a retrospective, observational, endline evaluation. The Lancet, 6:e500-e513.

Nutley, S. M., Walter, I., & Davies, H. T. O. (2008). Using Evidence: How Research Can Improve Public Services. Bristol: The Policy Press.

ODI (2016), Doing Development Differently: Who we are, What we're doing, What we're learning, ODI, London

Pawson, R. (2000) Middle-range realism, Archive Europeenne the Sociologie, Vol. XLI(2), 283-325

Pawson, R., and N. Tilley (2004) Realist Evaluation, paper prepared for the UK Cabinet Office, http://www.communitymatters.com.au/RE_chapter.pdf

Pritchett, L., Samji, S., and J. Hammer (2013), It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space, CGD Working Paper 322, CGD, London

Ramalingam, B., Jones, H., Tussaint, R., and J. Young (2008), Exploring the science of complexity: ideas and implications for development and humanitarian efforts, Working Paper 285, ODI, London.

Ravallion, M. (2015) The Economics of Poverty: History, Measurement, and Policy, Oxford University Press, New York

Sabet, S.M, and A.N., Brown (2018) Is impact evaluation still on the rise? The new trends in 2010-2015, Journal of Development Effectiveness, Vol. 10 (3), 291-304.

White, H. (forthcoming) Four waves of the evidence revolution. Nature (Palgrave Communications), Springer Press.

White, H, and E. Masset (2007) Assessing Interventions to Improve Child Nutrition: A Theory-Based Impact Evaluation of the Bangladesh Integrated Nutrition Project, Journal of International Development, 19, 627-652.

Wild, L., Booth, D., and C. Valters (2017), Putting theory into practice: How DFID is doing development differently, ODI Report, London

Contact us

Centre of Excellence for Development Impact and Learning
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org
@CEDIL2017