CDI Centre for Development Impact
PRACTICE PAPER
Innovation and learning in impact evaluation

Reality Bites: Making Realist Evaluation Useful in the Real World*

Abstract Realist evaluation asks 'how and why do interventions work or not work, for whom, and in what circumstances?' It holds promise as an approach that can help evaluate complex programmes, and provide nuanced insights to guide decisions about rolling out, scaling up, or trying out interventions elsewhere. This CDI Practice Paper, by Melanie Punton, Isabel Vogel, Jennifer Leavy, Charles Michaelis and Edward Boydell, presents lessons from four large, multi-country realist evaluations of complex interventions conducted by Itad since 2013. It argues that realist evaluation can add value by enhancing the clarity, depth, and portability of findings, helping evaluators deal with context and complexity in pragmatic ways, and providing helpful tools and lenses for implementers to critically appraise their programmes and generate learning. However, novice realist evaluators face a number of potential pitfalls, especially in large-scale evaluations. This paper shares lessons on how Itad has navigated these challenges, which may be helpful to others working in similar contexts in international development and beyond.

1 Introduction

The international development sector faces many thorny and interconnected global challenges, from climate change and cross-border trade to corruption and violent extremism. This has sparked a growing interest in evaluation approaches that can cope with complexity as well as generate lessons about how to tackle these kinds of challenges. Realist evaluation, developed by Pawson and Tilley in 1997, is one such approach. Through asking 'how and why do interventions work or not work, for whom, and in what circumstances?', it can provide robust insights and transferable lessons about why programmes succeed or fail, which can help

'A realist approach appealed as a way to evaluate an initiative working in challenging and complex international contexts. The realists' focus upon understanding "for whom" was particularly appealing in a development context, to build understanding of the distributional impacts of different types of interventions.' UK Department for Business, Energy and Industrial Strategy (BEIS), on why a realist approach was selected to evaluate ICF (pers. comm.) (see Box 1)

guide decisions about rolling out, scaling up, or trying out ideas elsewhere. This CDI Practice Paper discusses what makes realist evaluation distinct from other theory-based approaches and describes three principal benefits it brings, particularly to large evaluations of complex programmes in the international development sector.

The authors feel that a realist approach can add significant value where evaluations are commissioned to inform policies, programme decisions, and implementation practice. However, while it can help grapple with the challenges of evaluating the large-scale, complex, multi-country programmes that are typical in international development (see Box 1), realist evaluation poses a number of potential pitfalls for evaluators who are new to the approach. Through trial and error, we have learned several lessons that may be helpful to others working in similar contexts.

2 What is different about realist evaluation?

Realist evaluation is one of a family of theory-based evaluation approaches, alongside contribution analysis, Theory of Change-based approaches, and process tracing (Stern *et al.* 2012). Theory-based approaches are valuable because they *explain* why programmes work, rather than

Box 1 Four Itad realist evaluations[1]

Since 2013, Itad has worked on several large, multi-country realist evaluations of complex interventions across the global South. In 2018, the Realist Evaluation Learning Group was formed to discuss and reflect on Itad's experiences and distil what has been learned along the way.[2] This paper presents examples and lessons from four evaluations:

BCURE (2013–18). Itad led the evaluation of the Department for International Development (DFID)-funded, £13 million Building Capacity to Use Research Evidence (BCURE) programme, which worked within 11 African and Asian governments to build the capacity of civil servants and parliamentarians to use research evidence in decision making. The three-year evaluation spanned six countries in Africa and Asia, interviewing almost 600 stakeholders to explore the very different ways in which the six BCURE projects sought to build capacity – from training and mentoring, to network facilitation, to organisational support and institutional reform.

BRACED (2015–19). Itad was the Knowledge Manager for DFID's Building Resilience and Adapting to Climate Extremes and Disasters (BRACED) programme. BRACED aimed to build the resilience of more than five million vulnerable people against climate extremes and disasters, through a £110 million UK government-funded grant supporting more than 140 organisations in 15 consortia across 13 countries in East Africa, the Sahel, and Asia. Responsible for monitoring, evaluation, and learning, Itad used a realist synthesis approach to unpack how and why the projects did (and did not) work to improve resilience to climate extremes.

Compass (2015–20). Itad leads the Compass consortium, which provides monitoring, evaluation, and learning support for the UK's International Climate Finance (ICF). ICF is the UK government's commitment to help developing countries address climate change challenges through investing £5.8 billion between 2016 and 2021. The Compass consortium is conducting five strategic realist evaluations that address portfolio-level questions. Each evaluation involves a realist synthesis of secondary research drawing on over 300 ICF programmes in more than 50 countries, supplemented with realist primary research.

WAFM (2014–18). Itad led an evaluation of the West Africa Food Markets (WAFM) programme, which aimed to tackle the causes of the multiple market failures that exist in staple food markets in the region, in order to increase income and food security for producers and consumers. DFID invested £15 million between 2014 and 2019, with grants awarded to private sector companies in Burkina Faso, Ghana, Niger, and Nigeria. The evaluation incorporated a realist synthesis of six in-depth qualitative case studies of WAFM grants.

Source: Authors' own.

simply seeking to establish programme attribution – and 'when the aim is to learn so as to improve success or to replicate programmes elsewhere then explanations are needed' (Stern 2015). What sets realist evaluation apart from other theory-based approaches is its particular set of assumptions about programmes and the nature of reality, causality, and evidence, grounded in a realist philosophy of the world (Westhorp 2014). These provide the key to its rigour, explanatory power, and practical value. There are three distinctive features of a realist approach.

First, the realist understanding of 'programmes'.
Realist evaluators see programmes as, first and foremost, theories in action. These theories are often incomplete, implicit, or unconscious, and different stakeholders may hold very different theories about whether, how, or why a programme works. The consequence is that 'evaluation becomes a process of testing programme theories' (Pawson 2003). Realists also hold that programmes are inherently complex, and work through introducing new ideas or resources into existing social systems, in doing so changing the conditions that shape decision making and behaviour

(Pawson and Tilley 1997). Because programmes are 'complex interventions introduced into complex social systems' (Pawson 2013), no intervention works in the same way for everyone, all the time, but will have very different effects on different people in different contexts (Wong *et al.* 2016).

Second, the realist understanding of 'causality'.
Programme *outcomes* (any changes brought about by the programme) are understood as being caused by *mechanisms* (Pawson and Tilley 1997). Mechanisms in realist evaluation are not interventions. They are the – often invisible – forces, powers, processes, or interactions that lead to (or inhibit) change. They can be found in the choices, reasoning, and decisions that people make as a result of the resources a programme provides; the interactions between individuals or groups; and the powers and liabilities that things, people, or institutions have as a result of their position in a group or society (Pawson and Tilley 1997; Westhorp 2018). Mechanisms are 'triggered' when programme resources (e.g. information, money, expertise) interact with specific features of the *context* (individual, interpersonal, organisational, or institutional

Box 2 Context-mechanism-outcome configurations

Context-mechanism-outcome configurations (CMOs) are the core analytical building blocks of realist evaluation. They are variously described as propositions, hypotheses, or heuristics (Pawson and Manzano-Santaella 2012; Wong *et al.* 2013; Ravn 2019), and take the form of sentences or short paragraphs explaining how *mechanisms* interact with features of the *context* to generate *outcomes*. Both the BCURE and BRACED evaluations added intervention factors (I) to the configuration, to differentiate features of the intervention from features of the wider context, to create CIMOs. For example: 'In contexts where there are weak or non-existent market and institutional linkages (C), sequencing resilience-building activities appropriately and providing information in a timely manner (I), with people supported to apply new information (I), means that participants can make informed decisions about how they invest resources provided by the programme (M) in ways that are more likely to lead to resilient outcomes (O)' (BRACED final evaluation report).[3] For clarity, we will use the term 'CIMO' in this paper.

factors within the programme setting). A realist evaluation establishes a causal link between a programme and an observed outcome by homing in on these mechanisms, and developing and testing theories to explore them, in the form of context-mechanism-outcome configurations – see Box 2 (Befani 2012; Stern *et al.* 2012). In this way, realist evaluation uses theory to help us better understand reality.

Third, the realist understanding of 'knowledge' and 'evidence'. Given the complexity of the social world, and the limitations of human understanding and tools of enquiry, realist evaluators assume that theories can only ever be an approximation of reality (Williams 2018). Theory is therefore developed and tested in an iterative way, with the aim of reaching a refined theory that provides a good (enough) explanation of how and why outcomes have (and have not) occurred, while recognising that no theory can ever be fully and irrefutably confirmed as 'right' in a constantly shifting social world. Realist evaluation is methodologically eclectic, and so can use any relevant data collection or analysis tool and any relevant evidence to test theories (Marchal *et al.* 2012).

We argue that this distinctive realist approach to evaluation adds value in three ways:

1 **Enhancing clarity, depth, and portability of findings.** Realist evaluation encourages evaluators to develop clear and nuanced theories about how and why a programme works – forcing precision and a depth to the analysis

that may elude other theory-based evaluations. The focus on identifying mechanisms and exploring how they operate (or don't) in different contexts also allows evaluators to 'transfer' learning from one programme to another, which is particularly useful when commissioners wish to learn how to apply lessons from a programme elsewhere, scale it up or roll it out.

2 **Grappling with context and complexity, in pragmatic ways.** Realist evaluation recognises that, in the real world, programmes are messy, dynamic, and complex, analytical tools are imperfect, and evidence is incomplete. It provides tools and frameworks to help deal with these challenges through drawing evaluation boundaries, investigating the specific features of context that make a difference to interventions, and iteratively revising and testing theories in order to generate useful lessons.

3 **Engaging with stakeholders.** Realist evaluations do not necessarily require deep stakeholder engagement. However in our experience, involving the right people in the right ways at the right times is integral to making realist findings useful. When done well, realist evaluation can provide useful tools and lenses for stakeholders to critically appraise their programmes and deepen learning about how and why they influence change, helping to enhance the utility and credibility of findings.
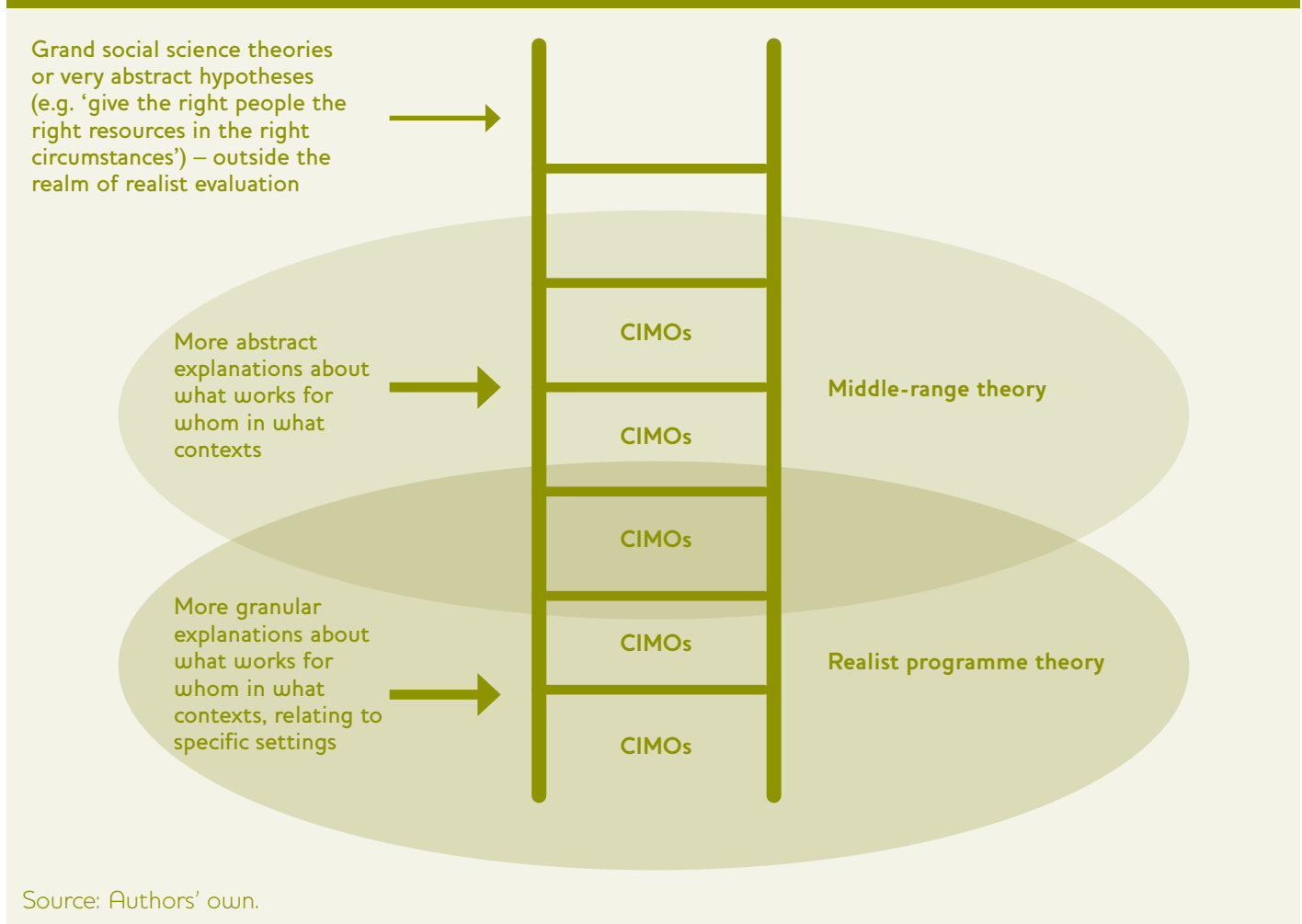
3 The value of realist evaluation

We now discuss each of these benefits in turn, give examples of how the four evaluations in Box 1 learned to harness them in practice, and describe the lessons we learned along the way.

Box 3 What is 'theory' in realist evaluation?

■ Realist **programme theory** explains (some of) 'how and why, in the "real world", a specific programme "works", for whom, to what extent and in which contexts' (Wong *et al.* 2016).

■ **Middle-range theories** lie between context-specific, granular working hypotheses about why things happen, and highly abstract **'grand theories'** of social behaviour and social change (Merton 1967; Pawson and Tilley 1997; Wong *et al.* 2013). They often provide an understanding of how *types* of intervention work in different *types* of circumstances.

■ **CMO (or CIMO) configurations** are hypotheses about how contexts (and intervention factors) interact with causal mechanisms to generate outcomes (see Box 2). They are used to unpack and interrogate both programme and middle-range theory.

Figure 1 **The 'ladder of abstraction'**

Grand social science theories or very abstract hypotheses (e.g. 'give the right people the right resources in the right circumstances') – outside the realm of realist evaluation

CIMOs

More abstract explanations about what works for whom in what contexts

CIMOs

Middle-range theory

CIMOs

More granular explanations about what works for whom in what contexts, relating to specific settings

CIMOs

Realist programme theory

CIMOs

Source: Authors' own.

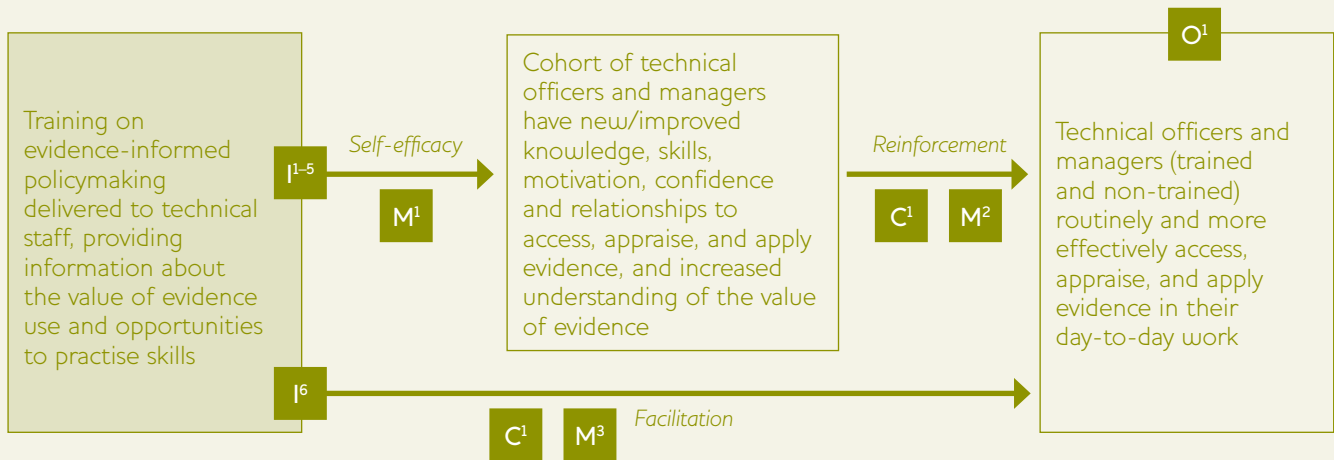### Enhancing clarity, depth, and portability of findings

Theory is at the heart of realist evaluation, and at the heart of what makes it useful. Three types of theory used by realist evaluators are outlined in Box 3.

The authors have found a 'ladder of abstraction' metaphor (Figure 1) useful to understand theory in realist evaluation (Pawson and Tilley 1997; Cartwright and Hardie 2012). This shows how CIMOs can be developed to explain specific activities in specific contexts at the bottom, and also provide more abstract middle-range explanations further up, while stopping short of engaging with grand, general theories at the top. The diagram also shows the overlap between programme theory and middle-range theory. Often, an evaluation will begin with very granular, context-specific programme theories at the bottom of the ladder, and eventually evolve these into more abstract programme theories later on, which apply across a wider range of settings.

Below, we look at three ways that theory in realist evaluation can bring clarity, depth and portability to evaluation findings.

Firstly, a realist programme theory is similar to a Theory of Change, a familiar concept in international development. Both generally start from 'stakeholder theory' – implicit and explicit ideas held by programme designers and participants, about why and how the programme works. However, a realist programme theory goes deeper than many Theories of Change, to hypothesise what goes on underneath the arrows that link outputs, outcome, and impact, and then test these hypotheses using CIMO configurations (Blamey and Mackenzie 2007). In our experience, this type of thinking is often not made explicit in evaluations, and even when it is (often in the form of 'assumptions' within the Theory of Change), it is not always systematically interrogated or evidenced – a problem also noted in the wider literature (Weiss 1997; Astbury and Leeuw 2010). In contrast, in realist evaluation the CIMOs form the heart of the investigation, bringing clarity and precision to causal linkages which may otherwise have remained fuzzy and unclear. The BCURE evaluation team found it helpful to use Theories of Change to provide a high-level overview of how the programme was intended to work, and then map CIMOs onto these to explain specific causal links. Figure 2 demonstrates how a CIMO configuration unpacks the connections between three boxes in a Theory of Change,

Figure 2 The theory underlying the arrows: an example from BCURE

Training on evidence-informed policymaking delivered to technical staff, providing information about the value of evidence use and opportunities to practise skills

I^{1-5} *Self-efficacy* M^1

Cohort of technical officers and managers have new/improved knowledge, skills, motivation, confidence and relationships to access, appraise, and apply evidence, and increased understanding of the value of evidence

Reinforcement C^1 M^2

O^1 Technical officers and managers (trained and non-trained) routinely and more effectively access, appraise, and apply evidence in their day-to-day work

I^6

C^1 M^3 *Facilitation*

CIMO 1: Where information is provided about the importance of evidence-informed policymaking and how to access, appraise, and apply evidence, alongside opportunities to practise skills, this can generate **self-efficacy** among technical officers, building their confidence in their ability to do their jobs or achieve goals (M^1) leading to individual behaviour change around evidence use (O^1). Behaviour change is more likely to be sustained where there are clear incentives (rewards, encouragement, reminders, audits or mandatory requirements) that motivate participants to apply their learning and **reinforce** changes in practice (M^2). This includes management support to encourage and provide space for participants to access, appraise, and apply evidence, which in turn depends on political incentives and a political environment that supports and promotes (non-symbolic) use of evidence as a means to improve policy (C^1). Behaviour change is also more likely where activities are closely targeted to individuals who can apply their learning because it is directly relevant to their day-to-day work (I^1), and where activities are practical and participatory (I^2), incorporate a focus on soft skills as well as technical skills (I^3), use knowledgeable, patient, and confident facilitators (I^4), tap into incentives to encourage participation (I^5), and provide practical tools, systems or processes (I^6) that **facilitate** trainees to do their jobs more easily or efficiently (M^3).

Source: Authors' own.

providing a detailed explanation of how, why, in which contexts, and through which intervention features BCURE led to civil servants using evidence more routinely in their day-to-day work. This CIMO is a third-generation theory, representing an evidenced hypothesis that had been iteratively tested and refined over three rounds of data collection across six countries.
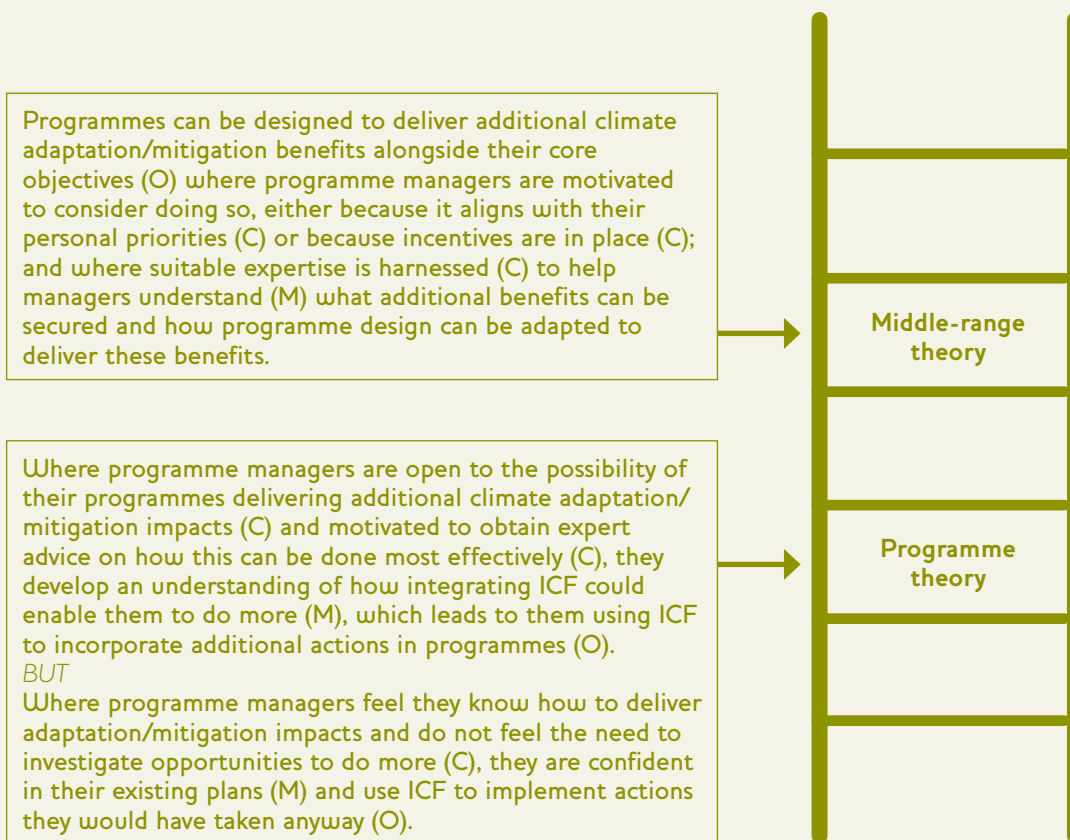
Secondly, realist evaluations encourage evaluators to 'stand on the shoulders of giants' by building on existing theory and evidence about the mechanisms through which a particular intervention works, in which contexts (Pawson 2013). Useful insights may also be available from quite different fields, recognising that similar causal mechanisms frequently underpin interventions across many sectors. For example, theories on social norms are relevant to a huge range of programmes seeking to change behaviour in areas ranging from corruption to contraceptive use (Mackie, Moneti and Denny 2015). Building on existing evidence and theory helps to develop richer insights that are grounded in what is already known, rather than starting from

scratch or relying exclusively on how stakeholders think their programme works (Pawson 2013; Astbury 2018). In many evaluations, this step is often missed or approached superficially due to time and resource constraints. Its central role in realist evaluation can help ensure theories and findings are both well grounded and useful to commissioners and implementers.

For example, there are a huge range of formal social science theories about how training and mentoring work. In order to understand how and in which contexts training courses led to behaviour change, the BCURE evaluation drew on adult learning theories such as the concept of 'self-efficacy', as well as theories on the 'diffusion of innovations' about how change can cascade throughout an organisation (Rogers 2003; Greenhalgh *et al.* 2004). This helped identify crucial contextual and intervention factors necessary to make training effective (see Figure 2). In the second year of the evaluation, the team noticed a pattern emerging: sustained changes in behaviour and practice were more likely where BCURE partners fostered a collaborative relationship with government and provided ongoing

Figure 3 **An example of moving between programme and middle-range theory, from the Compass evaluation**

Programmes can be designed to deliver additional climate adaptation/mitigation benefits alongside their core objectives (O) where programme managers are motivated to consider doing so, either because it aligns with their personal priorities (C) or because incentives are in place (C); and where suitable expertise is harnessed (C) to help managers understand (M) what additional benefits can be secured and how programme design can be adapted to deliver these benefits.

→ **Middle-range theory**

Where programme managers are open to the possibility of their programmes delivering additional climate adaptation/ mitigation impacts (C) and motivated to obtain expert advice on how this can be done most effectively (C), they develop an understanding of how integrating ICF could enable them to do more (M), which leads to them using ICF to incorporate additional actions in programmes (O).
BUT
Where programme managers feel they know how to deliver adaptation/mitigation impacts and do not feel the need to investigate opportunities to do more (C), they are confident in their existing plans (M) and use ICF to implement actions they would have taken anyway (O).

→ **Programme theory**

Source: Authors' own.

flexible and tailored support, rather than *ad hoc* capacity building through one-off interventions. By linking this insight to literature from the governance field on adaptive programming and 'development entrepreneurs' (Faustino and Booth 2014; Andrews, Pritchett and Woolcock 2016), we hypothesised the central importance of 'accompanying change, rather than imposing it'. This insight was further tested and confirmed in the final year of the evaluation and it became one of the key conclusions.

Finally, theory in realist evaluation helps make findings 'portable,' allowing us to apply knowledge about how a programme worked in one place to make decisions elsewhere (Pawson and Tilley 1997). Through uncovering and exploring the relationships between mechanisms and contexts, realist evaluations explain how *specific* activities work in *particular* settings. As evidence accumulates, and through linking programme findings to existing theories in the wider literature, it becomes possible to develop (or refute, or refine) CIMOs at the middle range that explain how *categories* of activities work in *types* of contexts. For example, the first Compass evaluation explored whether DFID programmes with integrated climate finance were designed to deliver additional adaptation or mitigation impacts and, if so, how and in what contexts.

CIMOs were developed and tested across 25 programmes to explore the circumstances in which integrating climate finance led to additional climate change benefits. This was then used to develop more abstract CIMOs at the middle range that aimed to apply across the large ICF portfolio (see Figure 3). The nuanced findings about how and why DFID programmes could be designed to deliver additional climate adaptation or mitigation impacts helped inform a review by the Independent Commission for Aid Impact (ICAI) of the UK's climate finance, and helped DFID identify where support could be provided to programme teams to help them understand climate risks and opportunities relevant to their programming.[4]

The BCURE final evaluation presented five tested CIMOs that helped explain how programmes can build capacity for evidence use, which have also proved 'portable' to new initiatives. For example, a World Health Organization-commissioned programme to improve the use of evidence in health decision making drew on insights about the importance of institutional support to help individuals apply new knowledge and skills (see Figure 2), and the value of accompanying decision makers through flexible, responsive support over time to ensure that interventions are appropriately situated within the local system.[5]

Box 4 Lessons on working with theory

Across all four evaluations discussed in this paper, one of the biggest challenges was understanding how to work with theory to develop useful and intuitive insights, without falling down a theoretical 'rabbit hole'. Through trial and error, we have identified several lessons that may help others conducting realist evaluations of complex programmes:

■ **Invest time to understand exactly what the intervention involves before developing detailed theories.** It is impossible to develop nuanced theories about how and why a programme works without a clear understanding of what the programme is doing in practice – what resources are being introduced to which stakeholders in what ways? Achieving this clarity is often far from simple in large and complex interventions with diverse packages of activities, where it can take time for implementers to nail down what their interventions will look like; where activities may evolve over time; and where the same language (e.g. 'training' or 'mentoring') can mask very different activities in different places. It is important in these cases to avoid developing theory too soon. For example, the WAFM evaluation team developed detailed theories across four thematic areas before grants had been awarded. However, the overwhelming majority of the awarded grants fell within only one of the four areas, making the other theories largely redundant.

■ **Do a literature review early on – and reserve resources to revisit it.** A substantial literature review (Punton *et al.* 2016) during the BCURE inception stage explored formal middle-range theories across a wide range of fields – from adult learning theory to organisational change – which were then integrated into the initial programme theory. However, as the evaluation evolved, new sources and fields of literature became relevant. By the third year, it was clear that the evaluation needed to draw more substantively on political science theories to unpack public sector incentives and blockages to evidence use, but by that time had limited time and resources to do so. In contrast, the BRACED programme evaluation benefited from a parallel research programme led by another partner in the Knowledge Manager consortium, and so was able to enrich the evaluation with evolving theory and research as the programme advanced.

■ **Work backwards from outcomes to find mechanisms.** We found it helpful to start by identifying tangible outcomes ('what changed?') and then work backwards ('how and why?') to find the mechanisms and contextual factors responsible. In BCURE, this approach helped keep the discussion focused and practical during semi-structured interviews, which in realist evaluation should involve explicitly discussing and refining theory with respondents (Pawson and Tilley 1997; Manzano 2014). However, this 'outcomes-first' approach can be difficult to apply early in the programme when outcomes are yet to emerge. It is also challenging with long-term, intangible outcomes like resilience, where it is not clear what success looks like at the outset. BRACED managed this through identifying and exploring short- and medium-term outcomes at the project mid-term, then looking at how they contributed to longer-term outcomes as the team learned more about what it means to build and strengthen resilience.

■ **Carefully balance formal theory and stakeholder theory.** There is a real risk that too much formal theory will make a realist evaluation seem overly academic to stakeholders; conversely, too much reliance on the informal theories of designers and implementers may miss important insights. It is the dialogue and interplay between these different kinds of theory that brings much of the value in realist evaluation (Astbury 2018). The BCURE team, as evaluators working in low- and middle-income countries, needed to make sure that it was grounding theories in stakeholders' experience rather than prioritising Western-centric theories. Practically, this meant that each phase of the evaluation started with a workshop to allow stakeholders to engage with the latest set of CIMOs and debate how they applied in their contexts – which both fed into the evolution of the team's programme theory, and supported implementers to engage with formal theory to enrich their work on the ground.

Grappling with context and complexity in pragmatic ways
Everyone agrees that 'context matters', and all evaluations need to take account of contextual factors that affect programme implementation and results. Evaluation commissioners are also increasingly interested in approaches that can evaluate programmes implemented in challenging environments and which exhibit features of complexity – such as multiple, evolving, and interacting components, implementation across several contexts, and feedback loops, tipping points, and emergent properties that cannot be predicted in advance (CECAN 2018). Context and complexity are closely related in realist evaluation, where complexity is viewed as an inherent feature of the always dynamic and unpredictable social contexts in which programmes operate (Pawson 2016; Coldwell 2019).

In many evaluations, context is treated as a 'macro variable' and relegated to a standalone section of the evaluation

Box 5 **Lessons on working with context and complexity**

Although CIMOs are a helpful framework to unpack context and deal with complexity, this is often no easy task. Our experience suggests some practical ways forward:

- **Use a programme theory to draw evaluation boundaries and set priorities.** In any intervention there are myriad mechanisms operating at different levels – for example, affecting individual behaviour, organisational culture, and institutional norms – which generate both short- and long-term outcomes. This is doubly true in large and complex programmes with multiple interlocking interventions. The Compass evaluations worked with stakeholders to identify the theories that were of greatest interest to policymakers. In contrast, the BCURE and WAFM evaluations initially attempted to look at everything, and as a result struggled to generate insights of sufficient depth. In the final year, the BCURE team used Theories of Change to guide discussions around where there were evidence gaps about how and why the programme worked, and where DFID was most interested in generating lessons for future programming. This was used to prioritise CIMOs to investigate in the final phase of the evaluation.

- **Collect data on both 'macro' and 'micro' contextual factors.** In the early stages of the BCURE evaluation, the team focused more on 'micro' features of context (e.g. around the characteristics of trainees) than on 'macro' features (e.g. around the nature of government systems and the influence of power, politics, and high-level incentives). This was because, as evaluators working with CIMO configurations for the first time, the 'micro' factors were generally easier to conceptualise and link to mechanisms. In the final year, a political economy analysis lens was built into the evaluation to address this gap. This provided a much better understanding of how the broader political context interacted with the mechanisms at work within the intervention. In future evaluations where macro-level political contextual factors are likely to affect programme mechanisms, a political economy exercise would be built in at the beginning.

- **View CIMOs as flexible framings, not rigid constructs.** The BCURE team has previously reflected that realist theorising risks becoming very linear: 'in this context, this intervention sparks that mechanism to lead to this outcome' (Punton, Vogel and Lloyd 2016). Forcing messy, complex processes into neat boxes can mean losing nuance in the process. Across the projects undertaken, the authors have found it beneficial to move away from treating CIMOs as rigid, simple formulae (C+M=O), and towards viewing them as flexible framings. It can help to start with broad theories that attempt to capture the essence of what might be causing change – in all its messy complexity – rather than agonising over whether something is a mechanism or a context and attempting to create neat configurations too soon. For example, the BRACED team began by developing broad 'if-then-because' statements – a tactic also recommended by Astbury (2018) – which were then refined into CIMOs *after* data collection. During the write-up stage, both the BRACED and BCURE evaluations also began presenting CIMOs as paragraphs rather than as simple sentences (see Figure 2 for an example). This provided much more explanatory freedom, and helped us depict the complex interplay of Is, Cs, and Ms influencing a particular outcome without over-simplifying.

Source: Authors' own.

report, rather than integrated into findings about how and why programmes succeed or fail (Ramalingam, Laric and Primrose 2014). However, in realist evaluation, the CIMO heuristic means that context remains central, and the interactions between context and mechanisms are systematically explored. Realist evaluation also helps evaluators avoid being overwhelmed by the enormity of potentially important contextual factors – recognising that, while infinite contextual factors may affect a programme, only some will be crucial to explaining why it works or does not work (Pawson and Tilley 2004). Interrogating a programme through the medium of CIMOs forces evaluators to home in on specific contextual factors that affect the mechanisms 'sparked' (or inhibited) by an intervention. An exploration of these factors can then be built into data collection; for example, through including

specific prompts in interview topic guides, and seeking out secondary sources that shed light on salient features of context, rather than asking generic questions about contextual enablers and barriers.

Realist evaluation also supports the exploration of change within and across multiple levels of a system. This is useful when evaluating large and complex programmes, with several components targeting different types of change (e.g. among individuals, communities, organisations and institutions, and policy). CIMO configurations help realist evaluators explore how programme activities and outcomes at these different levels are related, and how feedback loops operate between them (Westhorp 2012; Jagosh *et al.* 2015). For example, BRACED considered outcomes across four interconnected domains: individual and household level

change (related to resilient development outcomes); institutional change (from community to national and regional level); change to inclusion (including gender and other forms of social difference); and changes in the use of climate information. It looked at how these dimensions interacted, how they come together to contribute to resilience, and in turn how increased resilience can sustain each dimension into the future.

We also found that drawing on existing theory within the broader literature can be a route to bringing complexity thinking into evaluations. For example, the BCURE programme was initially designed based on fairly linear logic, namely 'if you build skills and give incentives, evidence use will increase and policy will improve'. The literature review identified more nuanced theories that viewed evidence-informed policymaking as a complex system, with change at multiple levels (individual, network, and organisational/institutional) required to make progress towards improvements in the quality of policy products and processes. This framing proved central to the evaluation, and CIMOs were developed to explore how change at one level (e.g. organisational) created conducive contexts for change at other levels (e.g. individual behaviour change), and vice versa. The framework was also useful to implementing partners who fed it into their programming (INASP 2016), and it influenced the design of a £17 million follow-up programme, where 'combining capacity interventions at different levels' was one of five key principles in the Business Case (DFID 2019).

Engaging with stakeholders

Realist evaluation is not necessarily a participatory approach, and it can be conducted with minimal engagement from programme implementers and evaluation commissioners. However, in our experience there is the potential to unlock considerable value if meaningful engagement is built into the design.

Firstly, as our theories represent imperfect and incomplete human attempts at explanation, and the social world is infinitely complex, it follows that any number of quite different theories might adequately explain how a programme works (Pawson 2013). However, only some of these theories will be useful and credible to the people the evaluation is intended to inform. We have found that co-creating theories collaboratively with implementers and commissioners helps evaluators focus on issues that are a priority for future programming, accommodate emerging priorities, and facilitate translation into operational recommendations that both policymakers and implementers trust. For example, the Compass evaluations have worked with an Evaluation Steering Group comprising policymakers with relevant experience from DFID, BEIS, and the Department for Environment, Food and Rural Affairs (Defra) along with evaluation specialists from the three departments. The steering group was closely involved

in theory development and revision through workshops at key stages, helping ensure the evaluations remained focused on areas of relevance to policymakers.

However, co-creating theory requires commitment and willingness on the part of stakeholders to engage with and interrogate their assumptions about how and why the programme is working, and careful facilitation from the evaluators to frame theory building and testing in practical ways, ensure all views are heard, and time engagement carefully. The WAFM evaluation team conducted theory-building workshops with implementers early on in the evaluation, to develop initial programme theories. There was little appetite to repeat this process when it became clearer what the interventions looked like, in order to revise or deepen the theories. Despite these challenges, the workshops helped the evaluation team understand the interventions in context, revealing that private sector predictability and flexibility were key factors in building trust among farmers so that they were more likely to adopt new practices and continue participating in interventions.

Secondly, realist evaluation is able to generate nuanced evidence that is practical and actionable. For implementers, this might involve granular insights about why a specific intervention worked in a particular place, in order to adjust and improve day-to-day programming. Commissioners generally require more abstract lessons about the types of interventions that work most effectively in different environments, to inform strategic decisions about what to fund in the future. By moving up and down the 'ladder of abstraction' (see Figure 1), realist evaluation can generate useful theories for the right stakeholder at the right time to support learning and adaptation.

For example, the BRACED implementing partners were able to use their realist mid-term evaluations as an opportunity to learn about what was working and why in their specific programme contexts. The power of the realist approach was that it allowed examination of how and why various packages of activities were either leading to expected outcomes or struggling to make progress – taking into account contextual, social, and political dynamics. These insights allowed implementing partners to have discussions with the fund manager about making necessary course corrections. At the same time, evaluators were able to synthesise evidence from across BRACED on how and why various packages of interventions contribute to greater resilience, generating evidence higher up the 'ladder of abstraction' that was of great interest to DFID. One spillover effect was that a number of implementers continued to apply a realist way of thinking to their annual reporting activities, providing depth and richer reflection to their routine monitoring reports by moving beyond listing activities and outputs, to considering what results were being achieved, for whom, and why change was or was not happening.

Box 6 **Lessons on working with commissioners and implementers**

■ **Develop tools and processes that support ongoing orientation, capacity building, and mentoring.** A realist approach requires flexibility and understanding around the need for iterative approaches to data collection and analysis in order to test evolving theory – potentially consulting new groups of stakeholders at different times or introducing new methods and tools as the evaluation progresses. It can take significant time to secure the buy-in and trust of commissioners for this, especially in multi-year evaluations where evaluation managers are likely to change. The BCURE evaluation had three evaluation managers in three years, requiring a substantial time investment to orientate new staff on the realist approach and win their confidence in the methodology. The team found it useful to develop short 'design papers' at strategic points in the evaluation to engage DFID, articulate the methodology, and orientate new staff. Where implementing partners are responsible for collecting realist data, capacity building is essential. The BRACED team developed guidance material and provided extensive one-on-one support with the project teams as they undertook their own realist evaluations. However, the team underestimated the time required to support implementing partners to 'think like realists'. Capacity building needs to be ongoing, and face-to-face time is helpful.

■ **Engage stakeholders in ways tailored to their needs.** The Compass evaluation helps manage the timing of stakeholder engagement through a formal Evaluation Steering Group, which produces terms of reference for each evaluation, advises on priorities, and participates in theory development and revision. The BCURE evaluation team used annual programme meetings to present and discuss emerging theories with implementing teams, as well as in-country workshops at the beginning of each round of data collection.

■ **Make theory practical.** In our experience, many stakeholders find the terminology of realist evaluation hard going, and the focus on theory building overly academic. Different types of stakeholders are likely to have differing levels of tolerance for engaging in theoretical discussions; for example, the WAFM programme struggled to engage private sector audiences in conversations about how and why they thought their business models worked. Realist evaluators need to find tools and language that can draw out the practical implications of theories for different audiences. We have found it helpful to park the language of CIMOs altogether when co-developing theory, instead taking a more intuitive approach of continually asking 'why' to dig for mechanisms and relevant contextual factors without introducing realist terminology.

■ **Keep the realist jargon 'under the bonnet' in reports.** Early BCURE reports structured findings around CIMOs, presenting them in lists and tables. However, this structure proved confusing for readers. In the final BCURE and BRACED reports, the team instead aimed to tell a coherent story about how and why the programmes had worked, with CIMOs displayed in pull-out boxes or in the form of 'vignettes' from specific projects throughout the report, rather than front and centre in the narrative. This proved more accessible and helped ensure the key messages of the evaluation were not buried beneath technical discussions about CIMOs.

Source: Authors' own.

However, evaluators need to be nimble in identifying key moments for consultation and critical reflection. In contrast to BRACED, the BCURE evaluation timeline was not well aligned with implementation decision points, and the contracting model emphasised the achievement of specific, predefined milestones, meaning projects had limited opportunity to adapt activities based on findings as theyemerged.

4 Conclusion: Realising the value of realist evaluation

This paper has reflected on the principles and practicalities of realist evaluation, drawing on lessons from four recent large-scale, multi-country evaluations. It has made the case

that a realist approach can bring value through enhancing the clarity, depth, and portability of evaluation findings, providing practical ways to grapple with context and complexity, and generating valuable learning opportunities for commissioners and implementers when meaningful stakeholder engagement is built in. Across these three areas, realist evaluation has enhanced the use and usefulness of the evaluations discussed in this paper, generating actionable findings that have fed into learning and decision making for programme implementers and funders.

However, understanding the assumptions behind realist evaluation and instilling them throughout evaluation design and practice requires time, effort, and an evaluation team willing to learn how to 'think like realists'. All four

evaluations discussed in this paper were conducted by teams new to realist evaluation, and we suspect that other novice realist evaluators may face similar challenges. The lessons in this paper can be summarised into the following advice:

■ **Set boundaries early.** Across the four evaluations, substantial time could have been saved by building in early processes to negotiate boundaries and set priorities with stakeholders, recognising that a realist evaluation cannot look at all dimensions of a complex programme and is best employed to investigate interventions and change processes that are least well understood. Where time and resources are in short supply, it may be more appropriate to use realist evaluation as one component within a larger evaluation investigating a few causal arrows or change processes that are particularly interesting or crucial.

■ **Keep it simple (at least at first).** Often, too much time was spent developing detailed theories too early, before the intervention was understood in sufficient depth. We suggest developing rough programme theories early on, rather than detailed CIMOs which may later prove to be irrelevant. Also, we feel it is beneficial to view CIMOs as a helpful heuristic, rather than a rigid framework, and to avoid realist jargon as much as possible when engaging with non-evaluators, and when writing up reports.

■ **Resource iteration and stakeholder engagement appropriately.** Underestimating the time required for evaluation is a common challenge, but the iterative

> '**The final independent evaluation report sets out very clear lessons and recommendations for DFID to take forward in future programmes that promote evidence-informed policymaking… We agree with all six recommendations, which have collectively fed into five key principles for our future… programming in DFID.**' DFID (2018)

nature of realist evaluation elevates this risk. The authors consistently underestimated how long it would take to revise theories, revisit the literature, understand and investigate emerging issues and priorities, build implementer capacity, and maintain positive relationships with new evaluation managers over the course of multi-year evaluations.

■ **Engage stakeholders at the right time in the right ways.** Timing is everything. Realist evaluation has the potential to be hugely valuable where evaluation activities can be aligned with programme decision points, while taking care to avoid 'theory fatigue' among busy programme staff.

With time, effort, and committed commissioners and implementers, realist evaluation can provide deeper understanding which can lead to better policy – it can be hard work, but it is worth it.

Endnotes

1 See https://itad.com/knowledge-and-resources/a-spotlight-on-realist-evaluation-in-itad/ for more information on each of the evaluations, including all published reports.

2 See https://itad.com/knowledge-and-resources/a-spotlight-on-realist-evaluation-in-itad/
3 This example is taken from the BRACED final evaluation report, available at: https://itad.com/reports/resilience-results-braced-final-evaluation-report/
4 See https://icai.independent.gov.uk/report/international-climate-finance/
5 Heightening Institutional Capacity for Government use of Health Research (HIGH-Res).

References

Andrews, M.; Pritchett, L. and Woolcock, M. (2016) *Doing Iterative and Adaptive Work*, CDI Working Paper, http://bsc.cid.harvard.edu/files/bsc/files/adaptive_work_cd_wp_313.pdf (accessed 8 February 2020)

Astbury, B. (2018) 'Making Claims Using Realist Methods', in N. Emmel *et al.* (eds), *Doing Realist Research*, London: Sage

Astbury, B. and Leeuw, F.L. (2010) 'Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation', *American Journal of Evaluation* 31.3: 363–81, https://doi.org/10.1177/1098214010371972 (accessed 20 February 2020)

Befani, B. (2012) *Models of Causality and Causal Inference*, London: Department for International Development,

www.ids.ac.uk/publications/models-of-causality-and-causal-inference/ (accessed 20 February 2020)

Blamey, A. and Mackenzie, M. (2007) 'Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges?', *Evaluation* 13.4: 439–55, https://doi.org/10.1177/1356389007082129 (accessed 20 February 2020)

Cartwright, N. and Hardie, J. (2012) *Evidence-Based Policy: A Practical Guide to Doing It Better*, New York NY: Oxford University Press

CECAN (2018) *Policy Evaluation for a Complex World*, Version 2.0, www.cecan.ac.uk/resources (accessed 20 February 2020)

Coldwell, M. (2019) 'Reconsidering Context: Six Underlying Features of Context to Improve Learning from

Evaluation', *Evaluation* 25.1: 99–117, https://doi.org/10.1177/1356389018803234 (accessed 20 February 2020)

DFID (2019) *Business Case: Strengthening the Use of Evidence for Development Impact (SEDI)*, https://devtracker.dfid.gov.uk/projects/GB-1-205283/documents (accessed 20 February 2020)

DFID (2018) *Management Response: Building Capacity to Use Research Evidence*, www.gov.uk/government/publications/building-capacity-to-use-research-evidence (accessed 20 February 2020)

Faustino, J. and Booth, D. (2014) *Development Entrepreneurship: How Donors and Leaders can Foster Institutional Change*, Working Politically in Practice Series, www.asiafoundation.org/resources/pdfs/DEODITAFWPIP2.pdf (accessed 8 February 2020)

Greenhalgh, T.; Robert, G.; MacFarlane, F.; Bate, P. and Kyriakidou, O. (2004) 'Diffusion of Innovations in Service Organizations: Systematic Review and Recommendations', *The Milbank Quarterly* 82.4: 581–629, https://doi.org/10.1111/j.0887-378X.2004.00325.x (accessed 8 February 2020)

INASP (2016) *Approaches to Developing Capacity for the Use of Evidence in Policy Making*, www.inasp.info/sites/default/files/2018-04/VY%20impact%20report.pdf (accessed 11 February 2020)

Jagosh, J. *et al.* (2015) 'A Realist Evaluation of Community-Based Participatory Research: Partnership Synergy, Trust Building and Related Ripple Effects', *BMC Public Health* 15.1: 725, https://doi.org/10.1186/s12889-015-1949-1 (accessed 8 February 2020)

Mackie, G.; Moneti, F. and Denny, E. (2015) *What are Social Norms? How are They Measured*, UNICEF/UCSD Center on Global Justice Working Paper, www.unicef.org/protection/files/4_09_30_Whole_What_are_Social_Norms.pdf (accessed 8 February 2020)

Manzano, A. (2016) 'The Craft of Interviewing in Realist Evaluation', *Evaluation* 22.3: 342–60, https://doi.org/10.1177/1356389016638615 (accessed 20 February 2020)

Marchal, B.; van Belle, S.; van Olmen, J.; Hoerée, T. and Kegels, G. (2012) 'Is Realist Evaluation Keeping its Promise? A Review of Published Empirical Studies in the Field of Health Systems Research', *Evaluation* 18.2: 192–212, https://doi.org/10.1177/1356389012442444 (accessed 20 February 2020)

Merton, R. (1967) *On Theoretical Sociology. Five Essays, Old and New*, New York NY: The Free Press

Pawson, R. (2016) 'The Ersatz Realism of Critical Realism: A Reply to Porter', *Evaluation* 22.1: 49–57, https://doi.org/10.1177/1356389015605206 (accessed 20 February 2020)

Pawson, R. (2013) *The Science of Evaluation: A Realist Manifesto*, London: Sage

Pawson, R. (2003) 'Nothing as Practical as a Good Theory', *Evaluation* 9.4: 471–90, https://doi.org/10.1177/135638900300900407 (accessed 20 February 2020)

Pawson, R. and Manzano-Santaella, A. (2012) 'A Realist Diagnostic Workshop', *Evaluation* 18.2: 176–91, https://doi.org/10.1177/1356389012440912 (accessed 20 February 2020)

Pawson, R. and Tilley, N. (2004) *Realist Evaluation*, www.communitymatters.com.au/RE_chapter.pdf (accessed 13 March 2014)

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, London: Sage

Punton, M.; Vogel, I. and Lloyd, R. (2016) *Reflections from a Realist Evaluation in Progress: Scaling Ladders and Stitching Theory*, CDI Practice Paper 18, Brighton: IDS, https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/11254 (accessed 11 February 2020)

Punton, M. *et al.* (2016) *How Can Capacity Development Promote Evidence-Informed Policy Making? Literature Review for the BCURE Programme*, Brighton: Itad

Ramalingam, B.; Laric, M. and Primrose, J. (2014) *From Best Practice to Best Fit: Understanding and Navigating Wicked Problems in International Development*, ODI Working Paper, www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9159.pdf (accessed 10 February 2020)

Ravn, R. (2019) 'Testing Mechanisms in Large-N Realistic Evaluations', *Evaluation* 25.2: 171–88, https://doi.org/10.1177/1356389019829164 (accessed 20 February 2020)

Rogers, E.M. (2003) *Diffusion of Innovations*, 5th ed., New York NY: Free Press

Stern, E. (2015) *Impact Evaluation. A Guide for Commissioners and Managers*, www.bond.org.uk/sites/default/files/resource-documents/impact_evaluation_guide_0515.pdf (accessed 10 February 2020)

Stern, E. *et al.* (2012) *Broadening the Range of Designs and Methods for Impact Evaluations*, DFID Working Paper 38, https://assets.publishing.service.gov.uk/media/57a08a6740f0b6497400059e/DFIDWorkingPaper38.pdf (accessed 10 February 2020)

Weiss, C.H. (1997) 'How Can Theory-Based Evaluation Make Greater Headway?', *Evaluation Review* 21.4: 501–24, https://doi.org/10.1177/0193841X9702100405 (accessed 20 February 2020)

Westhorp, G. (2018) 'Understanding Mechanisms in Realist Evaluation and Research', in N. Emmel *et al.* (eds), *Doing Realist Research*, London: Sage

Westhorp, G. (2014) *Realist Impact Evaluation: An Introduction*, www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9138.pdf (accessed 10 February 2020)

Westhorp, G. (2012) 'Using Complexity-Consistent Theory for Evaluating Complex Systems', *Evaluation* 18.4: 405–20, https://doi.org/10.1177/1356389012460963 (accessed 20 February 2020)

Williams, M. (2018) 'Making Up Mechanisms in Realist Research', in N. Emmel *et al.* (eds), *Doing Realist Research*, London: Sage

Wong, G. *et al.* (2016) 'RAMESES II Reporting Standards for Realist Evaluations', *BMC Medicine* 14.1: 96, https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-016-0643-1 (accessed 10 February 2020)

Wong, G.; Westhorp, G.; Pawson, R. and Greenhalgh, T. (2013) *Realist Synthesis: Rameses Training Materials*, London: Health Research Health Services and Delivery Research Program, www.ramesesproject.org/media/Realist_reviews_training_materials.pdf (accessed 10 February 2020)

> " Realist evaluation has enhanced the use and usefulness of the evaluations discussed in this paper, generating actionable findings that have fed into learning and decision making for programme implementers and funders. "

Centre for Development Impact (CDI)

The Centre is a collaboration between IDS (www.ids.ac.uk), Itad (www.itad.com) and the University of East Anglia (www.uea.ac.uk).

The Centre aims to contribute to innovation and excellence in the areas of impact assessment, evaluation and learning in development. The Centre's work is presently focused on:

(1) Exploring a broader range of evaluation designs and methods, and approaches to causal inference.
(2) Designing appropriate ways to assess the impact of complex interventions in challenging contexts.
(3) Better understanding the political dynamics and other factors in the evaluation process, including the use of evaluation evidence.

This CDI Practice Paper was written by **Melanie Punton**, **Isabel Vogel**, **Jennifer Leavy**, **Charles Michaelis** and **Edward Boydell**.