Article

Evaluation
2022, Vol. 28(1) 36–57

DOI: 10.1177/13563890211053028
journals.sagepub.com/home/evi
SAGE

Monitoring and evaluation for thinking and working politically

Thomas Aston
Independent Consultant, UK

Chris Roche
La Trobe University, Australia

Marta Schaaf
Independent Consultant, USA

Sue Cant
Independent Consultant, Australia

Abstract
This article explores the challenges of monitoring and evaluating politically informed and adaptive programmes in the international development field. We assess the strengths and weaknesses of some specific evaluation methodologies which have been suggested as particularly appropriate for these kinds of programmes based on scholarly literature and the practical experience of the authors in using them. We suggest that those methods which assume generative causality are particularly well suited to the task. We also conclude that factoring in the politics of uncertainty and evidence generation and use is particularly important in order to recognize and value diverse experiential knowledge, integrate understandings of the local context, accommodate adaptation and realistically grapple with the power relations which are inherent in evaluation processes.

Keywords
adaptive management, generative methods, participation, politics, rigour

Introduction

There is growing interest in politically informed and adaptive development programmes, but a recognition that the evidence base for this work is patchy (Laws and Marquette, 2018; McCulloch and Piron, 2019) and questions have been raised regarding the capacity of

Corresponding author:
Thomas Aston, Independent Consultant, 4 Kingswood Court, Hither Green Lane, London SE13 6TD, UK.
Email: thomasmtaston@gmail.com

conventional approaches to Monitoring, Evaluation and Learning (MEL) to adequately inform practice through actionable evidence (Pasanen and Barnett, 2019). Evaluators and development practitioners have wrestled with the challenges of evaluation use and usefulness for decades (Patton, 1978; Weiss, 1972). However, conventional approaches to MEL fail to take into account crucial issues of context and complexity (Chambers, 2015; Patton, 2010), which are integral to the dynamic character of programmes that aim to Think and Work Politically (TWP). Bridging the gap between these programmes and evaluation usefulness is, in part, a technical challenge, that is, one of methods, but it is also a political challenge (Roche and Kelly, 2012). In this article, we describe and address this dual challenge and explore the practice of trying to do so. We argue that programmes that seek to TWP assume a generative logic; a rigorous evaluation should therefore reflect this generative orientation. We also argue that MEL is an inherently political exercise, and therefore, practitioners need to be cognisant of this reality.

The technical challenges include recognizing the contested conceptual basis of much of the work that is described as TWP and the fact that it is usually highly contextually specific. Generative logic and causality based on open systems and taking context and power relations into account are central to these programmes in practice. This differs from experimental counterfactual logic, which is based on closed systems, controlling for context, and often disregards power (Greenhalgh and Manzano, 2021). This makes generating and applying generalizable knowledge difficult and potentially less useful than understanding and pooling local knowledge (Oliver et al., 2018) or producing knowledge that is transferable (Cartwright et al., 2020). Programmes that seek to TWP work from the premise that it is the combination or configuration of causes that lead to an outcome and it is the interaction of these causal and contextual factors which help to explain how and why outcomes are produced.

A forceful argument for generative causation in relation to adaptive programming should also prompt a much needed and broader debate about definitions of what constitutes rigour in MEL. This includes a greater emphasis on the importance of critical thinking, locally produced and grounded knowledge, the usefulness of evidence for programme adaptation and transferability linked to comparable contexts (Lincoln and Guba, 1985; Preskill and Lynn, 2016).

The political challenges arise from the fact that different stakeholders may well have divergent explicit and implicit personal, professional and learning agendas; value different forms of knowledge and have different power to insist on their preference. These tensions influence the forms of learning produced and the degree to which evaluators 'speak truth to power'.

We argue that the MEL of programmes that TWP needs to be 'technically sound and politically possible' (Faustino and Booth, 2014: 9). As such, this article builds on recent attempts to illustrate the kinds of MEL methodologies that are potentially a good fit for programmes that TWP (O'Keefe et al., 2014; Pasanen and Barnett, 2019) and adds to it by illustrating and reflecting on the more political dimensions of these processes – both in terms of choices regarding what and how evaluation is done, as well as whether findings are accepted – and how these interact with questions of methods and evidence. We do this by exploring several practical examples which have addressed some of these technical and political challenges together in practice.

This article makes a novel contribution in several ways. First, it explains the practical use of different generative approaches and makes a case for their use in understanding how development programmes that seek to TWP accomplish their goals. As well as assessing their

technical advantages and disadvantages, the article also analyses the degree to which these methodologies help address some of the inherent political dynamics of programme evaluation. Examples of real-life applications from the authors' own personal experiences are used to illustrate how these challenges can be addressed. In doing so, it raises important questions regarding definitions of rigour and hierarchies of knowledge that are of broader significance, as well as drawing out some of the implications derived from these cases and proposing some ways forward.

The challenges with doing MEL of TWP initiatives

The first challenge is that there is no single agreed way of TWP in development programmes. A TWP community of practice (CoP) emerged in acknowledgement of the failure of conventional, technical approaches to development assistance (Andrews, 2013; Booth and Unsworth, 2014). The CoP sets out three core principles: (1) strong political analysis, insight and understanding; (2) a detailed appreciation of, and response to, the local context and (3) flexibility and adaptability in programme design and implementation (Thinking and Working Politically [TWP] Community of Practice, 2015). Some suggest these principles have become 'the second orthodoxy' (Teskey, 2017), while others argue that there is confusion about what these terms mean, and variable commitment to the locally led nature of these initiatives (King, 2020).

Second, TWP's emphasis on context and on altering power relations entails a particular epistemic position. TWP work often engages in multi-scalar work, which in turn has feedback loops beyond the formal boundaries of programmes. So, conceiving context as a source of bias to be eliminated (Van Belle et al., 2016) and attempting to control for contextual features, as one would do through experimental approaches, set boundaries that are counterproductive to programming goals. A position that better fits the realities of TWP programmes is that the 'context within which a causal process occurs is, to a greater or lesser extent, intrinsically involved in that process' (Maxwell, 2004: 6). TWP thus requires a dynamic understanding of the relationship between context and mechanisms of change, which fits instead with a generative logic of causation (Falleti and Lynch, 2009; Greenhalgh and Manzano, 2021).

Third, there is little consensus as to the best ways of doing MEL for these kinds of programmes, although there is widespread agreement on the need for more appropriate theories of change and 'real-time' learning feedback loops to enable programme iteration and adaptation (Laws and Marquette, 2018; Teskey, 2017). Some proponents of TWP have called for higher standards of evidence based on experimental counterfactual logic and more comparative analysis (Dasandi et al., 2019; Laws and Marquette, 2018). However, as TWP programmes are adaptive by design, making the precise predictions that such approaches require is not only difficult but impedes these intended learning loops. Requirements for fixed programme logic in advance of implementation and concepts of 'fidelity' (Carroll et al., 2007), which demand interventions that are then delivered exactly as intended (Bauer et al., 2015), are also counterproductive. Moreover, proponents of counterfactual approaches within the TWP community have not teased out how these can address the challenges of measuring shifting institutional relationships and vested interests (Booth and Unsworth, 2014; McCulloch and Piron, 2019), or the fact that there may be multiple uncertain and interacting pathways of change and at different scales.

Table 1. Different ways to establish cause and effect in a reform programme.

Counterfactual logic	Generative logic	
To what extent did the intervention make a difference in pre-identified outcomes?	*How did a programme or portfolio of projects make a difference and through which combination of factors?*	*How did a particular project or intervention make a difference?*
Tells you if something 'on average' works, but not usually why or how. Can be focused on outcomes for particular groups.	Compares successful and less successful reforms to establish possible patterns of contextual and intervention factors which combine to produce more and less successful outcomes.	Explores the different possible explanations of a reform process to assess how they combine to produce an outcome.
Useful in helping to determine ultimate changes in people's lives, and, depending on the timeframe of the study, the sustainability of outcomes for particular reforms.	Critical in being able to develop more consistent practice based on comparative analysis, *as well as* programme and ongoing project-level learning.	Foundation of being able to explain how, in practice, a reform process was undertaken. Crucial as evidence *and* for ongoing project-level learning.
Methodologies: Randomized control trials (RCTs), quasi-experimental analysis, impact evaluation	Methodologies: Qualitative comparative analysis (QCA), realist evaluation	Methodologies: Process tracing, outcome harvesting

Source. Adapted from Schatz and Welle (2016).

Many critics of experimental counterfactual methods advocate for greater use of mixed methods to offset the aforementioned limitations (Deaton and Cartwright, 2018; Kabeer, 2019; Ravallion, 2018). However, this requires a greater understanding of the rationale for the mix of methods. It is the combination of these that provides evidence on not only *what* difference has been made but *how* this was achieved in a particular context, arguably *the* critical policy question for transfer purposes. This is a matter for generative logic, as Table 1 shows. We note that some authors such as Befani (2012) make a distinction between configurational logic (i.e. the identification of the combination of ingredients that explain an outcome) and generative logic (i.e. the processes by which these ingredients are combined – the recipe – to produce a given outcome). However, both rely on the quality of mechanistic within-case explanation, align with set theory about case membership and make asymmetric causal claims related to necessary and/or sufficient conditions (Beach and Pedersen, 2016). In this article, we, therefore, use the term *generative logic* to cover both.

It should be recognized that these two families of approaches (i.e. experimental counterfactual and generative) are each internally consistent but have different research cultures (Goertz and Mahoney, 2012). Goertz and Mahoney (2012: 42) distinguish between what they call 'effects of causes' approaches, which seek to assess average effects of particular variables, often favoured by quantitative scholars, and 'causes of effects' approaches, which are typically used in single cases to explain how outcomes or effects are produced by combinations of conditions or causal factors. Counterfactual logic, particularly as it is reflected in experimental designs, is chiefly concerned with the 'effects of causes', that is, average treatment effects, and pursues this through cross-case analysis in large-*N* studies. The focus is on a particular independent variable or the frequency of associations. In contrast, generative logic is chiefly concerned with the 'causes of effects', that is, necessary and/or sufficient conditions for a given

outcome, and pursues this primarily through within-case, and sometimes across-case, analysis in small-*N* studies. This kind of evaluative process is more like the work done by Sherlock Holmes[1] than that of randomized control trials (RCTs). At the same time, the use of generative logic through early theory development within RCTs for complex interventions is becoming more common and increasingly recommended (De Silva et al., 2014; Jamal et al., 2015), demonstrating the weakness of experimental methods without solid theory.

However, rigorous interpretation of results from evaluations based on generative logic requires careful reading and knowledge of the context and they are not easily summarized in brief highlights or infographics. Furthermore, it is often not recognized by those unfamiliar with generative analysis that it provides relevant evidence that other approaches cannot produce, namely, findings that are appropriate and helpful in assessing influence or contributions to policy reform and more transformational change (Junge et al., 2020).

Traditional MEL methods are largely designed to track progress in pre-planned projects where pathways of change are generally clear from the beginning and often represented in a logical framework. These methods have various requirements which depend on stringent conditions, such as that outcomes, counterfactuals and control groups can be clearly defined; there are identifiable primary causes and one or few primary effects; change is linear and a sample size is large enough for statistical analysis (Stern et al., 2012).

Yet, TWP programmes often have outcomes that are difficult to define precisely a priori; causes are typically made up of a combination of factors and similar outcomes can be produced by different causes. Furthermore, there may not be clear counterfactuals given the adaptation of strategies to context, change is often non-linear, control groups can be difficult or impossible to identify and infidelity of implementation may even be desirable for intended multi-scalar changes. Thus, most of the assumptions and stringent conditions required for experimental methods often do not hold and can be at odds with key principles of TWP. This does not mean that there is no place for appropriate counterfactual methods but that their use may be limited to specific parts of projects and questions. Nonetheless, the current focus on net effects and a preference for experimental designs, such as that displayed in a recent systematic review of participation, inclusion, transparency and accountability initiatives (Waddington et al., 2019), relegate generative methods and how and why questions to filling the gaps. This bias towards a specific approach and a particular hierarchy of knowledge may impoverish our understanding of TWP programming.

The MEL of politics and the politics of MEL

MEL for TWP is not technically straightforward; in addition, MEL is in itself an inherently political exercise (Eyben et al., 2015; Parkhurst, 2017). It is political because

- Different stakeholders have different views about what constitutes valid knowledge and evidence;
- These stakeholders have different degrees of power and therefore abilities to shape and indeed fund evaluations;
- Findings can affect the reputations, standings, careers and livelihoods of evaluators, as well as policymakers, agencies and their staff, and indeed communities; and
- Finally, evidence is not – and often should not be –the sole determinant of good decision-making – ethical and moral questions, equity issues and trade-offs between

> options require forms of deliberation which involve judgements – these are the stuff of small 'p' politics.

Failing to accommodate this reality can result in evaluations that are ignored or simply reinforce the status quo. There is little to no evidence showing that evidence *alone* shapes policy and practice. At the heart of this issue is the fact that different forms of knowledge are valued differently and this is often shaped by organizational practices. Those at the top of aid hierarchies and organizations tend to value unambiguous, succinct, often quantitatively based knowledge that seemingly offers a degree of scientific objectivity and rigour. In contrast, those on the front-line often need contextually specific, relational knowledge that helps them to navigate the messy, ambiguous reality and relationships that are required to make things happen (Honig and Gulrajani, 2018). Sector experts tend to value knowledge derived from the disciplinary or epistemic community to which they belong and from the methods which are most commonly used in that group, whereas practitioners and generalists might be less concerned with the disciplinary provenance of knowledge than the degree to which they can readily understand and use that knowledge in a practical manner.

This is not a new debate. Aristotle noted the difference between practical wisdom and theoretical reasoning and the importance of deliberation (Kinsella and Pitman, 2012). This distinction between generalized, decontextualized and 'thin' data and knowledge and contextual, embedded and 'thick' data and knowledge is important when it comes to determining what forms of knowledge might be needed to answer specific questions, or make decisions. However, a political take on this also recognizes that different actors in any situation will have varied ability and power to determine which forms of knowledge are likely to predominate. As such, as Wakefield and Koerppen (2017) and many other feminist scholars and practitioners note, MEL activities can both challenge and reinforce power relationships. This distinction and its implications are of particular relevance, given what has been described as the 'gender-blind' nature of much political economy analysis and associated TWP programmes (Derbyshire et al., 2018). Furthermore, once we accept that evaluation is inherently political, then it also means that evaluators must 'think and work politically' if their findings are actually going to make a difference.

In the light of these technical and political dimensions, in the next section, we explore some specific methodologies which have been suggested as particularly appropriate for the MEL of programmes which seek to TWP and attempt to draw out some lessons from both elements.

Best fit options

There are several methodologies that show promise; these can also be considered in combination. The methodologies proposed in this article are not new. Yet, they have only achieved serious recognition in international development circles in the past few years. Given our focus on practice, the article focuses on MEL methodologies with which the authors have had personal experience and fit with our understanding of TWP. It is thus illustrative of how methodologies might be used to undertake MEL of TWP and what challenges and benefits arise. We recognize that various other participatory and theory-based methodologies might also be a good fit (Pasanen and Barnett, 2019).

We use the categories proposed by Michael Crotty (1998) in discussing theoretical perspective, methodology and methods. Theoretical perspective is the philosophical stance that

Table 2. Politics of MEL.

Methodology	Summary	Technical features/ challenges	Evaluative and political context features/challenges
Realist evaluation/ synthesis	Generative and comparative methodology which develops hypotheses around the articulation of CMO statements that can be synthesized and aggregated into middle-range theories.	Focus is on 'what works, in which circumstances, and for whom'. CMO statements can appear more certain than intended. The approach can be theoretically and technically challenging and time-consuming.	Given its emphasis on hypothesis testing and its method neutrality, realist evaluation can be more palatable to decision makers who prefer positivist and/or 'quantitative' evaluations.
QCA	A rigorous qualitative methodology for comparative analysis of multiple cases in/ of complex settings/ interventions.	QCA combines case study and cross-case study analysis. QCA can be technically challenging. Supportive evidence is not always adequate, and verification is challenging.	The use of summary tables and the degree to which the presence or absence of key factors is demonstrably verifiable, including potential of algorithms to identify combinations of attributes in larger data sets, provides for a 'mixed methods' approach which is politically useful/ palatable.
Process tracing	Case-based and mechanistic method which uses evidence tests to assess inferential strength and compares alternative hypotheses.	Bayesian logic, evidence tests and rival hypothesis testing make inferences less vulnerable to bias and can be theoretically and technically challenging, and highly time-consuming.	If done in a participatory way, evaluation stakeholders can play a key role in defining which evidence is valued. Requires technical training which can impede ownership and participation at the beginning of the process.
Outcome harvesting	Participatory, actor-focused, within-case method. Works backwards from evidence of outcomes to assess the programme contribution.	Straightforward and easy-to-use method. Drafting quality outcome statements is crucial. However, risk of positive and confirmation biases due to limitations in triangulation.	As outcomes are typically drafted by programme participants, the method can be empowering. However, it relies heavily on evaluators' perspectives. Westernized focus on SMART reporting is a strength, but culturally limiting.

Note. Authors' construction. MEL = Monitoring, Evaluation and Learning; CMO = context, mechanism and outcome; QCA = qualitative comparative analysis; SMART = Specific, Measurable, Achieved, Relevant and Timely.

grounds methodologies, such as the experimental counterfactual and generative logics. Methodology is the strategy that links the choice of particular methods to the desired outcomes, such as realist evaluation, whereas methods are the techniques used to gather and analyse data, such as in-depth interviews or participant observation. We include both methodologies and methods in our discussion, all chosen for their particular relevance to TWP. Table 2 provides an overview of these methodologies and methods. We consider combinations of methods in the discussion section.

Realist evaluation

Realist evaluation is a type of theory-driven evaluation that asks, 'What works, in which circumstances, and for whom?', helping to answer questions of immediate relevance to the programme in question, as well as to refine theory (Pawson, 2013; Pawson and Tilley, 1997). There is some heterogeneity and debate among realist evaluators about both epistemological and methodological issues, but the description that follows largely reflects areas of consensus (Manzano, 2016; Marchal et al., 2012).

The realist evaluation understanding of generative causality is that an intervention works because actors make particular decisions in response to the intervention. In other words, actors' 'reasoning' changes in response to the resources or opportunities provided by the intervention. This combination of resources and reasoning comprises the mechanism that contributes to outcomes (Pawson and Manzano-Santaella, 2012). Mechanisms are at the heart of the 'black box' that generative approaches can open; they are not merely a restatement of programme activities, but the 'underlying entities, processes, or structures' that explain a causal relationship between the activities and the outcomes (Astbury and Leeuw, 2010; as cited in: Dalkin et al., 2015). Dalkin et al. (2015) argue that mechanisms can be activated by different degrees, depending on the context. This insight is particularly apt for programmes that aim to TWP, where human agency and relationships and local social and political dynamics play a key role, likely influencing the activation of a mechanism on a hyperlocal level.

Realist evaluations entail the elaboration of the underlying programme theory and then the collection of data to assess the realization of the theory. Context, mechanism and outcome configuration statements (or CMOs) are the building blocks of realist evaluations. These statements describe the causal regularities observed (Greenhalgh et al., 2009). Dalkin et al. (2015) suggest that MEL practitioners should differentiate between resources and reasoning in elaborating mechanisms, as this may help some evaluators to avoid confusing mechanisms and the context. The distinction between reasoning and resources might be especially pertinent in the context of TWP programmes, where changes in reasoning are germane to programme goals and trajectory. CMO statements are typically aggregated into more general middle-range theories or a refined programme theory. The approach provides an architecture for an evaluation, but it is formally method neutral, so a range of different quantitative and qualitative methods and tools may be used to collect the data (Kazi, 2003).

Realist inspired evaluations have been used to evaluate and understand programmes that explicitly or implicitly seek to TWP, primarily in the realm of social accountability initiatives. From a realist perspective, programmes and individuals exist in a larger social reality that is defined by interactions among individuals and institutions; causal mechanisms thus reside in individuals (agency) and in social relations (Marchal et al., 2012; Punton et al., 2020). It is precisely these things that TWP programmes aim to affect, as well as the broader social context for human and institutional action. For example, a realist evaluation of the 'Citizen, Voice and Action' (CVA) programme using social accountability to improve maternal and child health in Indonesia described the mechanisms that led to changed power relations within communities and between communities and health system authorities (Ball and Westhorp, 2018). While patriarchal norms typically limited women's input into community discussions, the programme under study explicitly facilitated their input through sex-segregated meetings. The evaluators concluded that the transparent process of sharing collective opinion with decision makers changed both the resources (information regarding collective opinion, particularly

women's opinions) available to decision makers and their reasoning (the incentives they faced to take action on the opinions). More specifically, the fact that the information represented a collective made it harder for decision makers to dismiss and the use of government standards 'legitimate[d] claims made by villagers and staff', providing 'authorisation for decision-makers to act' (Ball and Westhorp, 2018: 145).

There are some challenges that can arise in the application of realist evaluation to TWP, however. First, realist evaluation may not be user-friendly for key actors in TWP – grassroots actors (Lacouture et al., 2015; Marchal et al., 2012; Manzano, 2016; Pawson and Manzano-Santaella, 2012). Much of the foundational literature on realist evaluation is inaccessible to people without a significant background in evaluation, although participatory data collection can be used, and capacity building can be built into a realist evaluation. Relatedly, the realist understanding of context can be challenging for many to operationalize, especially when it comes to TWP. Even if MEL practitioners distinguish between reasoning and resources, as Dalkin et al. (2015) suggest, understanding the emergent properties that arise from the interaction between TWP interventions and context can still be challenging and defy neat distinctions between the two. Thinking of context as part of what engenders the outcome is feasible within a realist approach but can require access to significant data and analysis on the context, making it less feasible for less well-resourced evaluations.

Qualitative comparative analysis

Qualitative comparative analysis (QCA) is often seen to fit somewhere between large-*N* quantitative analysis and qualitative single-case analysis. Typically, it involves both within-case analysis and cross-case comparisons in order to explore 'constellations, configurations and conjunctures . . . where different conditions combine . . . to produce the same or similar outcomes' (Ragin, 2014: x). The methodology has its roots in political science and sociology and was seen by Charles Ragin, its main originator, as helping to put an end to the quantitative versus qualitative paradigm wars by drawing conclusions based on trends across cases while maintaining the strengths of rich case study analysis, including a holistic understanding of context and history (Ragin, 2014).

The Pacific Leadership Programme (PLP) supported a range of reform coalitions at a transnational Pacific level and in Samoa, Solomon Islands, Tonga and Vanuatu. This programme used QCA to identify factors (or combinations of factors) that were associated with coalitions achieving more or less success. Three sets of data were collected from PLP staff, the coalitions they supported and from programme documentation and combined for this analysis. The data set included 28 different coalitions, with 18 different outcomes of coalition 'success', and 65 different factors that may or may not have facilitated more successful interventions.

These raw data were then turned into what is known as a 'truth table' in QCA, that is, a matrix including all cases that records whether attributes were present or not present. In some instances, these data were 'crisp', that is, factors are 'present' or 'not present' in a clear and unambiguous verifiable way; in other cases, it was 'fuzzy', that is, judgements of differences in degree need to be turned into criteria for 'present' or not 'present' (Kraus et al., 2018). Using EvalC3, a predictive modelling Excel application (Davies, 2017), and human judgement, the data were analysed to identify attributes (and combinations of attributes) that were most strongly associated with 'more successful' or 'less successful' coalitions. This analysis was subsequently complemented with a social network analysis (SNA) to assess network

influence and size, in order to test one of the findings of the QCA exercise – a combination of methods that can be very useful in the analysis of policy reform (Fischer, 2011), and add a further level of credibility to the findings.

While the methodology did not find any definitive combinations of factors which explained more successful coalitions, it did help illustrate that there was no 'one size fits all' answer and there are different ways to achieve similar outcomes. However, there did seem to be a number of coalition attributes which, although they do not ensure success, may indicate whether a coalition with these attributes might be a 'better bet' than those without them. The analysis also indicated that predicting lack of progress can be as difficult as predicting relative success – that is, there is no one explanation for relative failure either. Finally, the analysis did indicate that certain country contexts were more propitious than others. The exercise illustrates how QCA, due to its cross-case focus, can add to an evidence base of how coalitional reform happens and contribute to the beginnings of a middle-range theory (Dasandi et al., 2019; Merton, 1968) which could be further developed and tested. This would be an important contribution in a field that has relied mostly on single-case studies.

QCA can be seen as technically complicated, requiring clear evidence of outcomes, as well as the inclusion of factors or conditions which are both objectively verifiable and definitively 'present' or 'not present'. This exercise shows how 'fuzzy-set' criteria generated through participatory methods can also be used to generate truth tables and it would be possible to undertake sensitivity analysis, that is, testing whether small changes in the scoring of inclusion criteria made a significant difference to the findings. Such flexibility is important for programmes whose outcomes are premised on relational or power shifts, which are hard to measure precisely.

For donors and those unfamiliar with generative causality, there is the risk of outcome evidence derived in this way as being seen as too 'fuzzy' and hard to verify objectively. On the contrary, one of the strengths of QCA is that the use of summary truth tables and visualizations across a range of projects can provide a succinct summary of a complex portfolio of projects for busy, time-poor managers. All of which suggest managing expectations of commissioners and clients about the strengths and weaknesses of the method is an important part of the exercise (Schatz and Welle, 2016).

Process tracing

Process tracing is a single-case, theory-based method. At the heart of the method is the idea of tracing causal mechanisms that link causes with their effects (i.e. outcomes) (Beach and Pedersen, 2019). Process tracing employs Bayesian logic, whereby new empirical evidence updates our confidence regarding the validity of hypotheses (Beach and Pedersen, 2019). Essentially, individual items of evidence are classified and appraised on the basis of their supposed inferential power, or 'probative value' (Befani and Stedman-Bryce, 2017). Evidence tests are commonly employed to test the strength of evidence at each step in a causal chain. For example, if one does not find the evidence necessary to the causal claim, the (hoop) test is failed, potentially invalidating the claim. However, in finding evidence that is unique to one's claim (smoking gun test), it is possible to confirm one's claim, potentially ruling out rival explanations (rival causal chains) within the case (doubly decisive test). The method is especially useful for evaluating relatively long and iterative causal chains over time (Naeve et al., 2017).

Both the Ghana Strengthening Accountability Mechanisms (GSAM) and the Journeys to Advancing Transparency Responsiveness and Accountability (JATRA) project in Bangladesh employed Bayesian process tracing (Befani and Stedman-Bryce, 2017). With the deliberate aim to promote learning, both projects conducted a 'partner-led' form of evaluation (Pasanen et al., 2018). An external evaluator worked with project teams to manage and coordinate data collection, analysis and reporting. They were supported by a semi-external quality assurer to help strengthen rigour.

Making a claim about an outcome a programme team believes it has influenced is highly political. In process tracing, evaluation stakeholders select observed outcomes ex post. When teams are enabled to choose their own outcome, there is typically an incentive to select the highest level of outcome they can to justify the high level of effort process tracing requires. In JATRA, this meant that the team evaluated a mechanism with six causal pathways and as many as 35 steps. With a partner-led approach, teams also had incentives to be light on considering rival explanations for changes they observed. So, teams need critical friends to help prompt critical thinking regarding what else contributed to or may explain the change they observed and counter confirmation bias.

Process tracing helps express extremely granular processes. Before the evaluation, the GSAM team had a single-pathway process map that described what *should* have influenced district assemblies' response to citizens' concerns about infrastructure investments. Through the process of developing casual chains, the team realized there were at least four different pathways to the same outcome. Getting diverse perspectives in the room to figure out what had actually happened and for different actors to explain their reasoning to others was extremely helpful in gaining consensus, as well as making causal chains more testable and robust (Aston, 2017).

In process tracing, what counts as credible evidence is context-specific. For instance, the JATRA evaluation team found that different social norms from non-governmental organizations, community-based organizations and government officials regarding the reliability of meeting records to track attendance meant that the probative value of this evidence was contextually relative. This prompted the project team to think more critically regarding the relative weight of evidence and to see the varying degrees of bias of different sources in a new light.

Having this grasp of probative value meant that project teams were also more efficient in data collection. In Bangladesh, of the 77 items of evidence identified for their causal chains, only half were required because the team recognized that some evidence was better at validating (or refuting) the project's contribution claim than others (Aston, 2018). As the GSAM team employed process tracing at midterm, the exercise streamlined the monitoring data requirements of partner organizations. The central team henceforth asked for fewer data of higher probative value in accordance with the causal mechanisms identified in the evaluation.

The key drawbacks of process tracing are that it is theoretically and technically complicated and requires teams to have a good understanding of their theory of change to develop testable and robust causal chains. Second, as a case-based methodology, process tracing is not ideally suited to assessing many different outcomes. It is helpful to achieve a depth of understanding, but not necessarily breadth. It thus requires teams to make potentially difficult and often political choices about which parts of their intervention merit a deeper dive and which do not.

Outcome harvesting

Outcome harvesting evolved out of outcome mapping and utilization-focused evaluation. It is participatory, actor-centred and focuses on contribution over attribution. In outcome harvesting, outcomes are defined as actors' behaviour changes such as actions, practices, relationships or policies (Wilson-Grau and Britt, 2012). It can be considered a generative method because it provides specific, in-depth, within-case explanations of how outcomes are produced in context. The Strengthening Advocacy and Civic Engagement (SACE) programme in Nigeria will be used to draw out salient aspects of the method which support TWP in practice.

Outcome harvesting works backward from what has been achieved to determine whether and how a programme contributed to the change. While progress marker data from outcome mapping can be used, process monitoring data are not strictly speaking necessary (Wilson-Grau and Britt, 2012). At the heart of the method is the development of short narratives (outcome statements) of who changed what, when and where, as well as a statement of significance and an explanation of how the programme contributed to the change. Outcome harvesting can be especially useful where outcomes are not easy to identify or measure a priori as it does not rely on indicators or milestones. This emphasis on emergent behaviour can be a strength and is one reason outcome harvesting has become so popular for complex programming in recent years.

The SACE programme in Nigeria worked through clusters of hundreds of organizations working on various thematic areas, supported by organizations that facilitated collaboration between clusters. Already employing outcome mapping and most significant change (MSC), SACE introduced outcome harvesting at an annual learning summit to help organization clusters see their individual work as contributions to an overarching goal as part of the midterm evaluation and to see their complementarities with other organizations also making contributions to that goal.

While some initiatives employ outcome harvesting simply as a final evaluation, as the SACE programme recognized, outcome harvesting can also be helpful in supporting teams to articulate significant changes as part of an annual review process. While sometimes challenging in practice, articulating outcomes as Specific, Measurable, Achieved, Relevant and Timely (SMART) is a key step in the method, and the SMARTer outcome statements are, the easier it is to confirm or refute contribution claims (Wilson-Grau and Britt, 2012). SACE teams developed outcome statements and presented these with evidence to cluster partners, alongside rubrics to assess the significance and the strength of evidence. Clusters also created a 'journey map' for outcomes to illustrate the change process to other clusters and project managers. This was then turned into an advocacy strategy matrix which served as a planning framework and outcome tracker, thereby contributing to programme adaptation directly.

Peer sense-making and critique were part of a wider effort in SACE to ensure that there was a common agenda among organizations, agreement on what success looks like, alignment of strategies and communication between stakeholders. This provided regular opportunities for partner organizations to ensure reasonable expectations, benchmark success collectively and to ensure that efforts would be more than a sum of their parts, illustrating connections between different contributions and outcomes. The process was also considered indispensable to trust and collaboration among cluster members (Root Change and Chemonics, 2018).

While its emphasis on emergent change is a strength, in some cases this creates incentives not to collect process-based monitoring data. This can potentially diminish initiatives'

capacity to make sense of actor-based interactions and adapt programming in a sufficiently timely manner. However, a greater weakness is the substantiation step (i.e. external corroboration) which suffers from issues of confirmation bias. This is the step which is most commonly skipped in practice (Smith, 2021) and this may stem from programmes harvesting more outcomes than it is realistic to substantiate.

Discussion

We have argued that TWP programming requires a dynamic understanding of the relationship between context and mechanisms, a looser approach to target setting to reflect the iterative nature of programming, recognition of multiple uncertain and interacting pathways of change at different scales and that it has a natural affinity for assessing 'causes of effects'. Programmes that seek to TWP assume a generative logic. Generative methodologies and methods are, thus, best suited to address their evaluative challenges.

The practical application of four best-fit methodologies/methods explained in this article revealed a number of technical advantages. First, we found that generative methodologies/methods help reveal valuable information on the inner workings of complex institutional environments, including otherwise unidentified connections between outcomes. Second, generative methods can lead to more trustworthy findings, including a more systematic exploration of rival explanations within cases. Third, we found practical dividends through the articulation of more testable and relevant outcomes ex-post, with more streamlined data collection and strategically useful findings. Together, they helped to shed light on causal mechanisms and on how contextual factors combine with programme attributes to contribute to outcomes and they can also support the development of middle-range theory. This, in turn, can help identify which programme attributes may be a 'better bet' in given contexts.

We also recognized that the methods/methodologies discussed in this article have different strengths and weaknesses in addressing the technical and political challenges of TWP programming. To synthesize and deepen our understanding, we now look at these strengths and weaknesses through the lens of rigour (Ton, 2012), acknowledging both the technical and political elements of rigour. Our account of the practical use of four best-fit methodologies/methods sheds light on the limits of context-independent understandings of validity (Maxwell, 2004), it raises questions regarding appropriate definitions of rigour and their alignment with the attributes of complex programmes (Stern et al., 2012) and it highlights the fundamental importance of responsiveness to evaluation stakeholders, as well as the politics of making evaluative judgements.

Bamberger et al. (2010: 6) argue that 'rigour is not determined solely by the use of a particular method as such, but rather the appropriateness of the "fit" between the nature of the problem being assessed and the particular methods deployed in response to it'. In this sense, we argue that not only should methodology/method choices reflect programme attributes, but so too should our definitions of rigour. Following Preskill and Lynn (2016), we see the following criteria as particularly relevant to TWP programming:

- *Reasoning*: Critical thinking is fundamental to evaluative reasoning and to thinking politically. This may include consideration of alternative explanations and interpretations, and a search for outliers (Schwandt, 2015; Scriven, 1977).
- *Credibility*: Credibility and the degree of confidence in findings speak to concerns of internal validity and the distinctiveness of effect patterns. The probative value of

evidence should be appraised in a contextually sensitive way (House, 1980; Lincoln and Guba, 1985; Scriven, 2008).

- *Responsiveness*: If programming aims to be locally led, then questions, methods and analyses should reflect local stakeholders' values and cultural context and be sensitive to their experiences and definitions of success, and evaluation criteria (Chambers, 2015; Patton, 2021; Stake, 2004).
- *Utilization*: If adaptation is key, then the quality of the learning process, actionable evidence and related utilization of evaluation findings are fundamental (Bamberger and Rugh, 2008; Julnes and Rog, 2009; Patton, 1978).
- *Transferability*: If context matters, then transferability and a reflection on potential moderating factors is more appropriate than generalizability. This relates to a more practice-oriented approach to external validity, with a greater emphasis on how the outcomes of an intervention are afforded by the context (Cartwright et al., 2020; Lincoln and Guba, 1985; Mark, 2011).

These criteria reflect TWP's principles of analysis, insight and understanding, response to local context, flexibility and adaptability. Particularly when TWP practitioners and evaluators are struggling to navigate wider methodological debates, these criteria might also help guide the appropriate choice of methodological combinations.

Combining qualitative and quantitative tools is one means to enhance rigour through methodological triangulation (Jimenez et al., 2018; Seawright, 2016). However, we would also suggest that combining different *parts* of generative methods through bricolage can also be an effective means to enhance rigour (Denzin and Lincoln, 1999). A number of the cases reviewed in this article combined parts of other generative methods in order to strengthen overall rigour. Both process tracing and realist evaluation are strong methodologies for reasoning and credibility because they are capable of developing precise mechanistic explanations (Stern et al., 2012; White and Phillips, 2012). Contribution analysis, for example, takes advantage of Bayesian reasoning and evidence tests within process tracing and has built these explicitly into its approach in some cases (Befani and Mayne, 2014; Ton et al., 2019).

QCA and realist evaluation are also strong methods to support transferability through their approach to comparison and development of middle-range theory. They can thus be used as an organizing frame for comparison, embedding aspects or steps of other methods such as process tracing tests. However, QCA, realist evaluation and process tracing are not intrinsically strong for responsiveness and utilization. Processes have to be consciously designed, or these methodologies/methods can be buttressed with elements of methods such as outcome harvesting that explicitly emphasize responsiveness and utilization.

Embracing the above criteria, we argue that responsiveness can also enhance rigour. In GSAM and JATRA, widening stakeholder inputs helped refine the programmes' causal logic, and in SACE, valuing the perspectives of different stakeholders helped to ensure that all significant outcomes were captured and that all partners were adequately represented. PLP, GSAM, JATRA and CVA all saw benefits from programme teams themselves gathering data. Peer sense-making and critique in SACE, GSAM, JATRA and CVA helped bolster the confidence of contribution claims and there were also wider benefits such as building trust in partners in SACE, GSAM and PLP. Furthermore, as programme staff were invested in the process and saw value in the data collected, it meant that they made use of evaluation findings to inform future strategy development in JATRA and SACE and to use evaluative reasoning

beyond the evaluation to support monitoring efforts in GSAM. Engaging communities, project/programme teams and partners in MEL can help bring out intangible processes around politics, changing power relations and collective action. Doing so can help provide real-time feedback to both the project/programme actors themselves and to those who are seeking to support them, allowing for learning, reflection and adaptation (Roche and Kelly, 2012).

As SACE, CVA and PLP all showed, the methodologies/methods discussed in this article can be combined with other methods to enhance rigour. These include MSC and the Bellwether approach, which can contribute to the responsiveness and utilization of an evaluation. MSC is explicitly designed to enhance responsiveness. By design, MSC's story-based approach valorizes the priorities of the programme's intended participants and beneficiaries, allowing stories of social and political change to emerge (Dart and Davies, 2003; Willetts and Crawford, 2007). MSC was helpful in both CVA and SACE to define outcomes as well as in mechanism development. Bellwether key informant interviews can also be used to incorporate the views of government officials, such as policymakers, making them apt to study interventions seeking to influence government policy or programmes (Coffman and Reed, 2007). The method entails using a baseline of key informant interviews with 'bellwethers' or influential people in the public realm at the beginning and the end of the programme. Including informants in the evaluation who are often the target of these initiatives also supports the credibility of findings by a form of substantiation similar to that in outcome harvesting, but with a greater degree of respondent independence.

This approach to rigour values participation while also problematizing platitudes in MEL regarding the importance of participation. Engaging stakeholders with differential power in the process of designing and undertaking an evaluation is not evident and some methods and methodologies can facilitate more equal footing by their design and/or in the way they are implemented. This remains far from straightforward, however. Even evaluation methods such as MSC, that are explicitly oriented to centre the individuals and communities' projects aim to empower, may not be able to overcome long-entrenched social hierarchies within communities. However, recognizing these challenges is essential to address them. Looking at rigour in the way we have defined it stakes a claim in the broader political economy of the development sector and the forms of knowledge and ways of knowing that are privileged.

These dimensions of rigour are also useful in exploring the political challenges of these different methodologies and methods, which we raised in 'The MEL of politics and the politics of MEL' section of this article. A significant challenge relates to issues of *reasoning*. In our experience, many decision makers and evaluation commissioners are unaware of the distinction between counterfactual and generative logics and the associated advantages and shortcomings of each. In part, this is arguably a hangover of older 'paradigm wars' related to purported 'gold standards' of evaluation, and 'hierarchies of evidence', and in part due to people's lack of familiarity with – often complex – methods that might be better able to capture non-linear, unpredictable and complex change. This, in turn, has an influence on the perceived *credibility* of approaches that often privilege key informants who are close to a given intervention in comparison with approaches that are seen to be more 'objective'. This is despite the fact that a *responsive* approach to evaluation would see this engagement with local stakeholders as an essential means of not only eliciting information that would otherwise be missed, but in shaping definitions of success, or failure. This engagement is also increasingly recognized as one of the key means by which evaluation findings are *utilized*. This is not simply because in doing so evaluators are helping policymakers or decision

makers, or practitioners to understand what evaluations conclude, but that their engagement often helps findings to be framed in ways more likely to resonate with their peers. At the same time, the interests of policymakers and decision makers also mean that accessing new evidence which is persuasively presented and which they feel they have helped co-produce can have political payoffs. This is why issues of *transferability* are also key. Knowing what worked is one thing. Knowing why it worked in a particular context, for whom and why it is unlikely to work somewhere else is much more useful and politically salient.

All of this suggests there is an important role in being better able to communicate in clear, simple language the value of generative methods in producing well-reasoned, credible and transferable findings, but also their potential for responsiveness and utility. This includes thinking about how divergent interests, power and politics shape not just what is evaluated, and how it is done, but the likelihood of uptake in the real world. Such approaches are key to exploring the relationships between interventions, systems and the contexts in which they emerge, and rendering evaluation better able to contribute to transformational change (Atkinson et al., 2021).

Conclusions

A decade ago, a narrative was gathering momentum in international development circles, which the UK Government coined simply as 'development is politics and politics is development'. The interest in the application of 'thinking and working politically' (TWP) has grown beyond programmes with more explicit links to politics and governance to broader recognition that all change has political dimensions, despite the fact that technical solutions still tend to dominate much development practice. Key TWP principles include strong political analysis, detailed appreciation and response to local cultural context and adaptive management. Employing these principles in practice can be technically and politically challenging for donors and programme implementers. Justifying the legitimacy of alternatives to counterfactual logic, as it is operationalized in experimental approaches, remains a challenge in a political economy of MEL which still often privilege linear approaches, premised on the assessment of the achievement of pre-identified outcomes.

This article argues that attempts to establish counterfactuals via experimental and quasi-experimental designs are of limited value to specific TWP questions and projects. Evaluation designs that support the explanation of how and why interventions work through the lens of generative causation are usually not only a better fit for the task at hand but also more useful for practitioners and policymakers. These methods can help to build and refine theory with respect to politics and governance, contributing to more transferable knowledge about TWP. We argued that being more explicit about generative causation prompts a much needed and broader debate about definitions of what constitutes rigour and how such definitions should respond to the goals and dimensions of programming itself (Preskill and Lynn, 2016).

Given the continued hegemony of counterfactual thinking, some evaluators and donors will continue to insist on their preferred approach (Aston, 2019). But just as TWP requires adaptability and diversity, so too does the MEL of TWP. A key challenge is to do this in ways that both use and capture adaptation, and which recognize plurality. At a time when debates on decolonizing development are merging with those on the politics of uncertainty (Scoones and Stirling, 2020), the importance of bringing multiple perspectives to bear on addressing common challenges has never been more salient. It is our hope that methodological discussions

such as this can help to elucidate pathways forward for MEL that recognize and centre experiential knowledge, integrate understandings of the local context, accommodate adaptation and realistically grapple with the politics and power relations that are inherent in the process. In other words, we hope that this article makes some contribution to the proliferating questions of *how* we change prevailing evaluation practice (Patton, 2021; Tyrrel et al., 2020). As Justin Parkhurst (2017) has noted, once we factor in politics and power, it becomes clear that one of the central issues we need to address is the governance of evidence generation and (mis-) use. Given that politics are inherent in interventions *and* evaluations, changes in MEL practices will also require changes in the governance of evidence. This is a heavy lift; we welcome further dialogue and debate in the MEL community about how to affect such change.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iDs

Thomas Aston https://orcid.org/0000-0002-4540-8363
Chris Roche https://orcid.org/0000-0001-9879-5769
Marta Schaaf https://orcid.org/0000-0002-7616-5966

Note

1. See Collier (2011) who uses Sherlock Holmes' investigation of the disappearance of the prize-winning horse Silver Blaze to explain process tracing.

References

Andrews M (2013) *The Limits of Institutional Reform in Development: Changing the Rules for Realistic Solutions*. New York: Cambridge University Press.

Aston T (2017) How to avoid toolsplaining: Thinking differently about social accountability. *CARE Insights*. Available at: https://insights.careinternational.org.uk/development-blog/how-to-avoid-toolsplaining-thinking-differently-about-social-accountability (accessed 25 November 2020).

Aston T (2018) Sounding clever or being smart? How to do more with less in evaluating governance programmes. *CARE Insights*. Available at: https://insights.careinternational.org.uk/development-blog/sounding-clever-or-being-smart-how-to-do-more-with-less-in-evaluating-governance-programmes (accessed 25 November 2020).

Aston T (2019) High priests, method police, and why it's time for a new conversation. *LinkedIn*. Available at: https://www.linkedin.com/pulse/high-priests-method-police-why-its-time-new-thomas-aston/ (accessed 25 November 2020).

Atkinson J, Lasbennes F and Nabarro D (2021) Reflecting on our times: Valuing transformative leadership in real-world 'living systems'. *American Journal of Evaluation* 42(1): 130–8.

Ball D and Westhorp G (2018) Citizen voice and action for government accountability and improved services: Maternal, newborn, infant and child health services. Final evaluation report, October. World Bank's Global Partnership for Social Accountability, Community Matters Pty, Washington, DC.

Bamberger M and Rugh J (2008) A framework for assessing the quality, conclusion validity and utility of evaluations: Experiences from international development and lessons for developed countries. *Paper presented at the American Evaluation Association conference*, Baltimore, MD, 7–10 November.

Bamberger M, Rao V and Woolcock M (2010) *Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development*. Washington, DC: World Bank.

Bauer MS, Damschroder L, Hagedorn H, et al. (2015) An introduction to implementation science for the non-specialist. *BMC Psychology* 3(1): 32.

Beach D and Pedersen RB (2016) *Causal Case Study Methods: Foundations and Guidelines for Comparing, Matching and Tracing*. Ann Arbor, MI: University of Michigan Press.

Beach D and Pedersen RB (2019) *Process-Tracing Methods: Foundations and Guidelines*, 2nd edn. Ann Arbor, MI: University of Michigan Press.

Befani B (2012) Modes of causality and causal inference. Broadening the range of designs and methods for impact evaluations: Report of a study commissioned by the Department for International Development. Department for International Development, London.

Befani B and Mayne J (2014) Process tracing and contribution analysis: A combined approach to generative causal inference for impact evaluation. *IDS Bulletin, Special Issue* 45(6): 17–36.

Befani B and Stedman-Bryce G (2017) Process tracing and Bayesian updating for impact evaluation. *Evaluation* 23(1): 42–60.

Booth D and Unsworth S (2014) *Politically Smart, Locally Led Development*. London: Overseas Development Institute.

Carroll C, Patterson M, Wood S, et al. (2007) A conceptual framework for implementation fidelity. *Implementation Science* 2: 40.

Cartwright N, Charlton L, Judean M, et al. (2020) Making predictions of programme success more reliable. CEDIL methods working paper, Centre for Excellence for Development Impact and Learning (CEDIL), Oxford.

Chambers R (2015) Inclusive rigour for complexity. *Journal of Development Effectiveness* 7(5): 327–35.

Coffman J and Reed E (2007) Unique methods in advocacy evaluation. Harvard family research project. Available at: http://www.pointk.org/resources/files/Unique_Methods_Brief.pdf

Collier D (2011) Understanding process tracing. *Political Science and Politics* 44(4): 823–30.

Crotty M (1998) *The Foundations of Social Research: Meaning and Perspective in the Research Process*. London: SAGE.

Dalkin S, Greenhalgh J, Jones D, et al. (2015) What's in a mechanism? Development of a key concept in realist evaluation. *Implementation Science* 10: 49.

Dart J and Davies R (2003) A dialogical, story-based evaluation tool: The most significant change technique. *American Journal of Evaluation* 24(2): 137–55.

Dasandi N, Laws E, Marquette H, et al. (2019) *What does the evidence tell us about 'thinking and working politically' in development assistance?* WIDER working paper 2019/12. New York: UNU – WIDER.

Davies R (2017) EvalC3. Available at: www.evalc3.net

De Silva MJ, Breuer E, Lee L, et al. (2014) Theory of change: A theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials* 15: 267.

Deaton A and Cartwright N (2018) Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* 210: 2–21.

Denzin NK and Lincoln YS (Eds) (1999) *The SAGE Handbook of Qualitative Research*, 3rd edn. Thousand Oaks, CA: SAGE.

Derbyshire H, Siow O, Gibson S, et al. (2018) *From Silos to Synergy: Learning from Politically Informed, Gender Aware Programs – Developmental Leadership Program*. Birmingham: University of Birmingham.

Eyben R, Guijit I, Roche C, et al. (2015) *The Politics of Evidence and Results in International Development: Playing the Game to Change the Rules?* Rugby: Practical Action Publishing.

Falleti TG and Lynch JF (2009) Context and causal mechanisms in political analysis. *Comparative Political Studies* 42(9): 1143–66.

Faustino J and Booth D (2014) *Development Entrepreneurship: How Donors and Leaders can Foster Institutional Change*. San Francisco, CA: Asia Foundation.

Fischer M (2011) Social network analysis and qualitative comparative analysis: Their mutual benefit for the explanation of policy network structures. *Methodological Innovations Online* 6(2): 27–51.

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

Greenhalgh J and Manzano A (2021) Understanding 'context' in realist evaluation and synthesis. *International Journal of Social Research Methodology* 24(3): 1–13.

Greenhalgh T, Humphrey C, Hughes J, et al. (2009) How do you modernize a health service? A realist evaluation of whole-scale transformation in London. *The Milbank Quarterly* 87(2): 391–416.

Honig D and Gulrajani N (2018) Making good on donors' desire to do development differently. *Third World Quarterly* 39(1): 68–84.

House E (1980) *Evaluating with Validity*. Beverly Hills, CA: SAGE.

Jamal F, Fletcher A, Shackleton N, et al. (2015) The three stages of building and testing mid-level theories in a realist RCT: A theoretical and methodological case-example. *Trials* 16: 466.

Jimenez E, Waddington H, Goel N, et al. (2018) *Mixing and Matching: Using Qualitative Methods to Improve Quantitative Impact Evaluations (IEs) and Systematic Reviews (SRs) of Development Outcomes CEDIL Inception Paper 5*. London: CEDIL.

Julnes G and Rog D (2009) Evaluation methods for producing actionable evidence. In: Christie CA, Mark MM and Donaldson SI (eds) *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Thousand Oaks, CA: SAGE, 96–134.

Junge K, Cullen J and Iacopini G (2020) Using contribution analysis to evaluate large-scale, transformation change processes. *Evaluation* 26(2): 227–45.

Kabeer N (2019) Randomized control trials and qualitative evaluations of a multifaceted programme for women in extreme poverty: Empirical findings and methodological reflections. *Journal of Human Development and Capabilities* 202: 197–217.

Kazi M (2003) Realist evaluation for practice. *British Journal of Social Work* 33(6): 803–18.

King M (2020) Why does local agency matter? Enabling the policy space for aid recipients. *Socarxiv* 6: ew8.

Kinsella EA and Pitman A (2012) Professional practice and education. In: Kinsella EA and Pitman A (eds) *Phronesis as Professional Knowledge: Practical Wisdom in the Professions*. Rotterdam: Sense Publishers, 1–11.

Kraus S, Ribeiro -Soriano D and Schüssler M (2018) Fuzzy-set qualitative comparative analysis (fsQCA) in entrepreneurship and innovation research: The rise of a method. *International Entrepreneurship and Management Journal* 14(1): 15–33.

Lacouture A, Breton E, Guichard A, et al. (2015) The concept of mechanism from a realist approach: A scoping review to facilitate its operationalization in public health program evaluation. *Implementation Science* 153: 10.

Laws E and Marquette H (2018) *Thinking and Working Politically: Reviewing the Evidence on the Integration of Politics into Development Practice over the Past Decade, Working and Discussion Paper*. Birmingham: TWP Community of Practice, University of Birmingham.

Lincoln Y and Guba E (1985) *Naturalistic Inquiry*. Newbury Park, CA: SAGE.

McCulloch N and Piron LH (2019) Thinking and working politically: Learning from practice – Overview to special issue. *Development Policy Review* 37: O1–O15.

Manzano A (2016) The craft of interviewing in realist evaluation. *Evaluation* 22(3): 342–60.

Marchal B, Van Belle S, Van Olmen J, et al. (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.

Mark MM (2011) New (and old) directions for validity concerning generalizability. In: Chen HT, Donaldson SI and Mark M (eds) *Advancing Validity in Outcome Evaluation Theory and Practice New Directions in Evaluation, 130*. New York: John Wiley & Sons, 31–42.

Maxwell JA (2004) Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher* 33(2): 3–11.

Merton RK (1968) *Social Theory and Social Structure*. Glencoe, IL: The Free Press.

Naeve K, Fischer-Mackey J, Puri J, et al. (2017) *Evaluating advocacy: An exploration of evidence and tools to understand what works and why*. 3Ie working paper 29. New Delhi, India: International Initiative for Impact Evaluation (Ie) 3.

O'Keefe M, Sidel JT, Marquette H, et al. (2014) *Using action research and learning for politically informed programming*. Research paper 29, Developmental Leadership Program, University of Birmingham, Birmingham.

Oliver S, Roche C, Stewart R, et al. (2018) *Stakeholder engagement for development impact evaluation and evidence synthesis*. Centre of Excellence for Development, Impact and Learning, inception paper 3. London: Centre of Excellence for Development, Impact and Learning.

Parkhurst J (2017) *The Politics of Evidence: From Evidence-based Policy to the Good Governance of Evidence*. Abingdon: Routledge.

Pasanen T and Barnett I (2019) *Supporting adaptive management: Monitoring and evaluation tools and approaches*. Overseas Development Institute (ODI) working paper 569. London: ODI.

Pasanen T, Raetz S, Young J, et al. (2018) *Partner-led evaluation for policy research programmes: A thought piece on the KNOWFOR programme evaluation*. Working paper 527. London: ODI.

Patton MQ (1978) *Utilization-Focused Evaluation*. Beverly Hills, CA: SAGE.

Patton MQ (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford Press.

Patton MQ (2021) Evaluation criteria for evaluating transformation: Implications for the coronavirus pandemic and the global climate emergency. *American Journal of Evaluation* 42(1): 53–89.

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

Pawson R and Manzano-Santaella A (2012) A realist diagnostic workshop. *Evaluation* 18(2): 176–91.

Pawson R and Tilley N (1997) *Realistic Evaluation*. Thousand Oaks, CA: London: SAGE.

Preskill H and Lynn J (2016) Redefining rigour: Describing quality evaluation in complex adaptive settings. *FSG Reimagining Social Change*. Available at: https://www.fsg.org/blog/redefining-rigor-describing-quality-evaluation-complex-adaptive-settings (accessed 25 November 2020).

Punton M, Vogel I, Leavy J, et al. (2020) *Reality bites: Making realist evaluation useful in the real world*. Centre for Development Impact practice paper no. 22. Brighton: Centre for Development Impact.

Ragin C (2014) *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*, 4th edn. Berkeley, CA: University of California Press.

Ravallion M (2018) *Should the randomistas (continue to) rule?* CGD working paper 492. Washington, DC: Center for Global Development.

Roche C and Kelly L (2012) *The evaluation of politics and the politics of evaluation*. Background paper 11. Birmingham: Developmental Leadership Program, University of Birmingham.

Root Change and Chemonics (2018) Volume 3: Participatory monitoring, evaluation, and learning in complex adaptive environments strengthening advocacy and civic engagement (SACE) program in Nigeria. *Report*, December. Root Change and Chemonics, Washington, DC.

Schatz F and Welle K (2016) *Qualitative comparative analysis: A valuable approach to add to the evaluator's toolbox? Lessons from recent applications*. Centre for Development Impact practice paper number 13. Brighton: Centre for Development Impact.

Schwandt T (2015) *Evaluation Foundations Revisited: Cultivating a Life and Mind for Practice*. Stanford, CA: Stanford University Press.

Scoones I and Stirling A (2020) *The Politics of Uncertainty: Challenges of Transformation*. London: Routledge.

Scriven M (1977) *Reasoning*. New York: McGraw-Hill.

Scriven M (2008) A summative evaluation of RCT methodology: An alternative approach to causal research. *Journal of Multidisciplinary Evaluation* 5(9): 11–24.

Seawright J (2016) *Multi-method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.

Smith R (2021) Use of outcome harvesting for monitoring in dialogue and dissent alliances: Findings from a survey and discussions. Available at: https://www.outcomemapping.ca/resource/use-of-outcome-harvesting-for-monitoring-in-dialogue-and-dissent-alliances-findings-from-a-survey-and-discussions

Stake R (2004) *Standards-based and Responsive Evaluation*. Thousand Oaks, CA: SAGE.

Stern E, Stame N, Mayne J, et al. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations: Report of a Study Commissioned by the Department for International Development*. London: Department for International Development.

Teskey G (2017) *Thinking and working politically: Are we seeing the emergence of a second orthodoxy?* Governance working paper series no. 1. ABT Associates Governance and Development working paper series, Australia.

Thinking and Working Politically (TWP) Community of Practice (2015) *The Case for Thinking and Working Politically: The Implications of 'Development Differently'*. Birmingham: TWP Community of Practice, University of Birmingham.

Ton G (2012) The mixing of methods: A three-step process for improving rigour in impact evaluations. *Evaluation* 18(1): 5–25.

Ton G, Mayne J, Delahais T, et al. (2019) Contribution analysis and estimating the size of effects: Can we reconcile the possible with the impossible? Centre for Development Impact practice paper. Brighton: Centre for Development Impact.

Tyrrel L, Kelly L, Roche C, et al. (2020) *Uncertainty and COVID-19: A turning point for monitoring evaluation, research and learning*. ABT ASSOCIATES Governance and Development working paper series number 7. ABT Associates, Australia.

Van Belle S, Wong G, Westhorp G, et al. (2016) Can 'realist' randomised controlled trials be genuinely realist? *Trials* 17(10): 1–6.

Waddington H, Sonnenfeld A, Finetti J, et al. (2019) Citizen engagement in public services in low- and middle-income countries: A mixed-methods systematic review of participation, inclusion, transparency and accountability (PITA) initiatives. *Campbell Systematic Reviews* 2019: E151025.

Wakefield S and Koerppen D (2017) Applying feminist principles to program monitoring, evaluation, accountability and learning. *Oxfam*. Available at: https://policy-practice.oxfam.org/resources/applying-feminist-principles-to-program-monitoring-evaluation-accountability-an-620318/

Weiss C (1972) *Evaluating Action Programs*. Boston, MA: Allyn & Bacon.

White H and Phillips D (2012) Addressing Attribution of Cause and Effect in Small N Impact Evaluations: Towards an Integrated Framework. Technical report 15 International Initiative for Impact Evaluation working papers. New Delhi, India: International Initiative for Impact Evaluation.

Willetts J and Crawford P (2007) The most significant lessons about the most significant change technique. *Development in Practice* 17(3): 367–79.

Wilson-Grau R and Britt H (2012) Outcome harvesting. Outcome Mapping Learning Community. Available at: https://usaidlearninglab.org/sites/default/files/resource/files/outome_harvesting_brief_final_2012-05-2-1.pdf

Thomas Aston is an Independent Consultant. He has a particular interest in participatory and theory-based approaches to evaluation, chiefly in social accountability and governance programmes.

Chris Roche is Director of the Institute for Human Security and Social Change at La Trobe University. He is interested in how attempts to promote social change is assessed.

Marta Schaaf is an Independent Consultant. She has a particular focus on accountability processes in sexual and reproductive rights programmes.

Sue Cant is an Independent Consultant and PhD Student at Charles Darwin University. Her main area of interest is in social accountability and governance programmes.