Process Tracing as a Practical Evaluation Method: Comparative Learning from Six Evaluations

Alix Wadeson, Bernardo Monzani and Tom Aston
March 2020

Table of Contents

Acronyms and Abbreviations 2

i) Introduction 3

iii) Process Tracing, Bayesianism and Inferential Strength 6

iv) The Evaluation Cases 8

a. Ghana's Strengthening Accountability Mechanisms 9

b. Journeys to Advancing Transparency Responsiveness and Accountability 10

c. CARE International's Cocoa Life Project in Côte d'Ivoire 11

d. IIED's support to the Least Developed Countries Group 12

e. Government of Canada's commitment to girls' education in crisis contexts 13

f. Oxfam America and Climate Change and Energy Advocacy 14

v) What Have We Learned? 15

a. Participation 15

b. Theories of Change 17

c. Methodological decisions 18

d. Mitigating bias 23

vi. Recommendations to improve practice and use 26

vii. Bibliography 29

Table of Figures

Figure 1: Illustrative Example of a Causal Mechanism 5

Figure 2. Variation in Evaluation Approach 9

Figure 3. Levels of Participation in Evaluation 15

Figure 4. Causal Complexity across Evaluations 19

Acronyms and Abbreviations

CAP	Community Action Plan
CARE	Cooperative Assistance for Relief Everywhere
CCAT	Climate Change and Energy Advocacy Team (at Oxfam America)
CDCOM	Community Development Committees
CMO	Context-Mechanism-Outcome configuration
COP	Conference of Parties
CSO	Civil Society Organization
DM&E	Design Monitoring & Evaluation
FGD	Focus Group Discussion
G7	Group of Seven
G20	Group of Twenty
G7CSO	Group of Seven Civil Society Organizations
GoC	Government of Canada
GPSA	Global Partnership for Social Accountability
GSAM	Ghana Strengthening Accountability Mechanisms
IIED	The International Institute for Environment and Development
JATRA	Journeys to Advancing Transparency Responsiveness and Accountability
KII	Key Informant Interview
LDC	Least Developed Countries
OI	Oxfam International Confederation
OUS	Oxfam America
QuIP	Qualitative Impact Protocol
VoIPT	Veil of Ignorance Process Tracing
ToC	Theory of Change
UNFCCC	United Nations Framework Convention on Climate Change
UP	Union Parishad
WVC	World Vision Canada

i) Introduction

One important current trend in evaluation discourse amongst international development practitioners is an interest in finding appropriate methods for evaluating the impact of interventions that Buffardi, Pasanen, and Hearn (2017) refer to as the "hard to measure." This includes issues of 'abstract concepts,' 'multi-dimensional problems,' 'equifinality' and 'multifinality' in pathways of change (see also, ODI, 2018). Examples of "hard to measure" changes include efforts to shift gender norms and empower women; advocacy for pro-poor government policy and budgeting; and improving governance. For interventions with such goals, purely quantitative approaches to evaluation and simply assessing performance against logical framework model indicators are inadequate and fail to meet the challenge of evidencing how and why change happens. Accordingly, organizations are starting to apply more fit-for-purpose approaches such as Process Tracing, Contribution Analysis, Outcome Mapping and Outcome Harvesting (see Pasanen and Barnett, 2019). This further reflects increasing recognition of Theory of Change (ToC) as a key foundation to test assumptions and better understand the interplay of complex dynamics and relationships amongst stakeholders in a given change process or system (Vogel, 2012).

Despite the exploration of such methods, there is still a relative dearth of examples of practical learning and evidence of good practices in applying these approaches, including Process Tracing, to help inform the broader international development sector. In order to support better practice and contribute to the evidence base, this paper presents comparative learning from the evaluation of six international development initiatives that applied various forms of Process Tracing. While these initiatives span across diverse contexts and pursued different aims, they are connected by a common thread: all six case studies centre around efforts to influence others - often decision makers and those in power - around aspects such as practices of consultation and inclusion; public policy; and resource allocation.

The paper is organized in the following manner. We first explain Process Tracing and review common definitions. Secondly, we consider the potential value added of an explicitly Bayesian approach to Process Tracing. Next, we discuss the six cases where Process Tracing was applied, noting similarities and differences. Then, we explore key practical learning emerging from the cases and insights from the use of different forms of Process Tracing across different programming contexts. These reflections are organized under four meta-themes of participation, Theory of Change, methodological decisions, and mitigating bias. Finally, we present our key recommendations, ending with practical tips, targeted at practitioners and evaluators interested in applying Process Tracing, especially for initiatives falling under the 'influencing' umbrella.

ii) What is Process Tracing?

Process Tracing has been referred to as a method (Beach and Pedersen, 2013; Collier, 2011), a tool (Bennett, 2010; Collier, 2011) and a technique (Bennett and Checkel, 2014) for data collection and analysis (Beach, 2016). Whether considered method, tool or technique, its use in evaluation is relatively recent (Stedman-Bryce, 2013; Punton and Welle, 2015; Befani and Stedman-Bryce, 2016; Befani *et al.* 2016). Process Tracing is considered particularly useful for the evaluation of interventions based on ToCs, for example, governance and advocacy initiatives, which are difficult to evaluate with experimental and statistical methods (Befani and Mayne, 2014; Neave *et al.* 2017; Stedman-Bryce *et al.* 2017).

Process Tracing is commonly referred to as a case study methodology (Stern *et al.* 2012; Beach, 2016), and its approach to causation is generative (theory-based). Therefore, at the heart of Process Tracing is the idea of tracing causal mechanisms that link cause "X" with its effect "Y" (i.e. outcome) (Beach, 2016; Beach and Pedersen, 2019). The starting point is

provided by an observable outcome; a theory is then defined and broken down in a series of causal mechanisms, which together are deemed logically necessary and sufficient for achieving the outcome.[1] The causal mechanisms allow for the identification of relevant and appropriate evidence. This evidence is then put through four different tests, which refer to different forms of probative value (see Box 1). Based on this value, the theory can be validated, or not, and contribution to impact effectively assessed.

Box 1. Similarities and differences among the four process tracing tests

Straw-in-the-Wind (neither confirmatory nor disconfirmatory): If the evidence is observed, this is not sufficient to confirm the hypothesis. If the evidence is not observed, this is not sufficient to reject the hypothesis.
Hoop Test (disconfirmatory): If the evidence is not observed, the hypothesis is rejected. If the evidence is observed, the hypothesis is not rejected (it 'goes through the hoop', passes the test); but it is not confirmed, either.
Smoking Gun (confirmatory): If the evidence is observed, the hypothesis is confirmed. If the evidence is not observed, the hypothesis is not confirmed; but it is not rejected, either.
Doubly Decisive (both confirmatory and disconfirmatory): If the evidence is observed, the hypothesis is confirmed. If the evidence is not observed, the hypothesis is rejected.

Collier, 2011: 825; Befani and Stedman-Bryce, 2016: 4

Process Tracing is part of a family of so-called theory-based approaches to evaluation, whose benefits have been recognized even by some advocates of experimental methods, who underscore the importance of understanding causal mechanisms and local context (Bates and Glennerster, 2017; Gugerty and Karlan, 2018). However, as Derick Beach (2016) notes, there remains considerable discord regarding the definition of a causal mechanism. One important review found over 24 definitions of causal mechanisms proposed by sociologists, political scientists, and philosophers of science in the last 40 years (Mahoney, 2001 in Falleti and Lynch, 2009). In Process Tracing itself, there is also no clear consensus. Beach and Pedersen (2019: 31) refer to 4 general types:

1. **Descriptive narratives** of events between cause and outcome, or intervening events (Mahoney, 2012: 571, Mahoney, 2016).
2. **Intervening variables**, that is, factors intervening between cause and outcome (King, Keohane, and Verba, 1994: 87; Gerring, 2007).
3. **Minimalist mechanisms**, that is, causal process observations (CPOs) or "diagnostic evidence" assumed to be linked to empirical fingerprints (George and Bennett, 2005: 6; Brady and Collier, 2011; Bennett and Checkel, 2014).
4. **Mechanisms as systems** of interlocking parts transmitting causal forces between cause and outcome (Beach and Pedersen, 2019).

Beach and Pedersen (2019) argue that mechanisms should be understood as system of interlocking parts that transmits causal forces between a cause (or a set of causes) and an outcome (Bhaskar 1978; Beach and Pedersen 2013, 2019: 38). In this reading, we should not be looking for causes and outcomes, or simply providing a descriptive narrative of events which logically follow in a temporal sequence (George and Bennett 2005; Bennett and Checkel, 2014; Mahoney, 2016), [2] but rather explaining the process, the causal links, and the relationships between events. Causal mechanisms are found *in between* X and Y (Falleti and Lynch, 2009; Beach and Pedersen, 2019: 2 – 3, 31 – 32). A mechanism is not, in other words, about 'nuts and bolts,' but rather 'cogs and wheels' (Hernes, 1998: 78, in Beach, 2016: 465)

[1] In fact, process tracing does not require necessity or sufficiency. The only requirement of a mechanism to be causal is that it transfers some form of causal forces from C to O (Beach and Pedersen, 2016: 36).
[2] There can also be feedback loops, as various iterations may be required for transmission along the causal chain.

and how causal forces are transferred from cause to outcome. It is about how the wheels move in the process; indeed, it is as much about *how* things happen as what is happening (Falleti and Lynch, 2009).

While this may appear challenging conceptually, it is more straightforward empirically than looser definitions of mechanisms, because it works at a lower level of analytical abstraction. In this form, mechanisms consist of entities (actors, organizations, etc.) which are the forces engaged in activities, while activities are the producers of change, which transmit causal forces (Beach and Pedersen, 2013, 2019: 4). Similar to Realist Evaluation (see Pawson and Tilly, 1997; Pawson, 2013), this form of Process Tracing needs to clearly explain the contextual conditions which must be present, the mechanism which links cause to outcome, and outcome (CMO). Figure 1 below provides a simplified relevant example.

Figure 1: Illustrative Example of a Causal Mechanism

<table>
<tr><th>Cause</th><th>Activities</th><th>Part 1</th><th>Part 2</th><th>Part 3</th><th>Outcome</th></tr>
<tr><td>Advocacy to influence international development policy to be more gender transformative.</td><td>(verb) and Entities (noun)</td><td>Civil society actors (entity) identify policy window of opportunity and coordinate on a joint case for a gender-transformative policy commitment (activity).</td><td>Civil society actors (entity) provide key evidence (activity) to public servants (entity) on the needs for and benefits of gender-responsive development programming.</td><td>Public servants (entity) engage with high-level policy makers (entity) and promote civil society advocacy messaging and evidence (activity).</td><td>The government announces new feminist policy commitment across its funded international development initiatives.</td></tr>
<tr><td colspan="6">Contextual conditions that must be present</td></tr>
</table>

Adapted from Beach and Pedersen, 2019: 71

While there is surprisingly no cross-referencing (amongst literature reviewed or experts consulted), Process Tracing has various similarities with Realist Evaluation, which is also about how given certain contextual conditions and causal mechanisms lead to outcomes. Realist Evaluation underlines that it is not interventions that create change; *people* create change (Pawson and Tilly, 1997). Mechanisms are about people's choices; their volition, bounded by a particular social context. Process Tracing is typically focused on observational data. However, people's motivations are not necessarily observable – they do not necessarily leave traces or fingerprints. Going beyond descriptive narratives of events should entail an assessment not merely of how such events are linked but *why* an actors' actions would trigger a particular cognitive and emotional responses in another actor. It requires not only an 'if, then' statement, but rather an 'if, then, *because'* statement. Without being explicit about actors' motivations, it is quite possible to misidentify mechanisms. This is not to say that Realist Evaluation itself is invariably better at causal identification, but that a more explicit effort to reflect on actors' reasoning should make identification easier and the logic clearer to those examining the casual mechanisms.

Colin Hay (2016: 500) has argued that Process Tracing is a 'laudable ambition but not a methodology'. For Hay, identifying processes is hard enough, let alone tracing them. He thus suggests Process Tracing often labels an ambition, assuming that the processes have been correctly identified in the first place. However, this is a criticism that can be made of any method which seeks to assess changes which have a high degree of causal complexity. Arguably, a concern with an agent's "reasoning" which we find in Realist Evaluation may help (Pawson and Tilly, 1997; Pawson, 2013), but as Stern *et al.* (2012) note, for complex programs [and processes], 'the best that can be expected is plausible interpretation rather than firm

"proof"' (2012: 34). Taking the reasoning of Sherlock Holmes, Collier, Brady, and Seawright (2010) argue that it is useful to think of Process Tracing as a search for "clues" which help us compare rival explanations.

Notwithstanding, Process Tracing does offer the opportunity to assess how firm this proof might be through the application of evidence tests. This allows one to see how far general causal indications stand up to scrutiny, to help confirm a specific explanation and to reject rival explanations. A common way to describe Process Tracing is to see it as akin to the work of a detective investigating a crime or a lawyer presenting evidence to a jury. Similarly, Process Tracing requires the evaluator to (i) make predictions about what empirical evidence would be left by an intervention if the hypothesis were true, considering evidence we would both expect and hope to find and assess its probative value; (ii) gather empirical evidence; and (iii) assess whether we can trust the evidence found for each part of the mechanism. At this point, one can make a judgement regarding whether hypothesized mechanisms are present or not (Beach and Pedersen, 2019: 4, 178). In Process Tracing, it is possible to have causal claims, which are deemed necessary (i.e. needed to pass hoop tests), complementary claims, which play a supportive rather than necessary role, and rival claims (i.e. smoking gun or doubly decisive evidence found for *rival* claim), which either diminish confidence in or would rule out the clam under study.

Process Tracing's use of four evidence tests, as described in Box 1 above, is one of its distinctive features. Of these, hoop tests and smoking gun tests are generally the most useful to grade and prioritize evidence that can either confirm or disconfirm a hypothesis about a causal claim. A hoop test is useful to disconfirm a hypothesis, but, because of its low uniqueness, it is not enough to confirm a hypothesis. A smoking gun test, for its part, has high uniqueness and is sufficient to confirm the hypothesis (Punton and Welle, 2015); failing this test does not, however, disconfirm a hypothesis. In practice, the four tests are useful to identify and prioritize evidence, based on whether this is deemed necessary and unique, and to assess, in a rigorous manner, its "probative value", weighting it in a way that helps us to discriminate between rival explanations (Befani and Stedman-Bryce, 2016; Fairfield and Charman, 2017).

As international development interventions tend to have relatively clear parameters, often expressed in a logical framework (flowing from inputs, to outputs, to outcomes), Hay's (2016) criticism regarding misidentification is less problematic for project evaluation than for research, and only becomes a potential concern at higher levels of a causal chain, and determining the significance of contribution. However, at the same time, there is a greater threat of confirmation bias; over-focusing on data that fit a particular hypothesis and overlooking data that undermine it (Fairfield and Charman, 2018). Given this, as White and Philips (2012) note, more effort is required to ensure that small-n evaluations minimize the biases which are likely to arise from the collection, analysis, and reporting of predominantly qualitative data. One way to do this is through Bayesianism.

iii) Process Tracing, Bayesianism and Inferential Strength

In recent years, we have seen increased interest in moving from analogies to the formal application of Bayesian logic to Process Tracing (Bennett, 2008; Bennett, 2014; Schmitt and Beach, 2015; Humphreys and Jacobs 2015; Befani and Stedman-Bryce, 2016; Fairfield and Charman, 2017, 2018; Beach and Pedersen, 2019). Fairfield and Charman (2017) view the turn to Bayesianism as a watershed for in-depth, small-n research.

Bayesian reasoning is a means of updating our views about which hypothesis best explains the phenomena or outcomes of interest as we learn additional information (Fairfield and Charman, 2018: 6). The logic of inference in Process Tracing is Bayesian, in that new

empirical evidence updates our confidence regarding the validity of theories (or hypotheses), and this updating depends upon how unique this empirical evidence is to the hypothesis (Beach and Pedersen, 2013). Essentially, different pieces of evidence are classified and graded on the basis of their supposed inferential power or "probative value."

Contribution Tracing marks the most explicit attempt to use Bayesian logic in Process Tracing and to quantify confidence that an intervention has contributed to an outcome for evaluation.[3] In Contribution Tracing, Bayesian updating is used to assign a (prior) probability (how likely it is) that the various components of your contribution claim exist; and ultimately whether your claim holds true. It is common to have a 50:50 (no information) prior. Equal priors avoid biasing the initial assessment in favour of any particular hypothesis (see Fairfield and Charman, 2017 and Box 2 below). You assign a probability for each item of evidence you identify, and then gather your prioritized data with the best probabilities. We update our confidence in a contribution claim (posterior), as we observe (or do not observe) items of evidence, comparing the claim against rival explanations.

Fairfield and Charman (2017) view classification of tests as unnecessary within a Bayesian framework, since evidentiary confirmation is always a matter of degree, not type, and inference is always governed by the logic of Bayes' rule. They suggest that the main aim is to compare rival hypothesis, and what matters in this exercise is "inhabiting the world of each hypothesis". You ask how surprising (low probability) or expected (high probability) the evidence would be in that world. While it is possible to conduct rigorous Process Tracing without formal tests, the tests are helpful to determine the probative value of evidence. Simply put, most available evidence is likely to help you show that what you did happened as you say you did (hoop tests) rather than whether there is a unique connection between what you did and what caused change (smoking guns).

As Fairfield and Charman (2017) note, in social science, inferences are commonly based on the accumulated weight of evidence from many clues, none of which is strongly decisive. However, the accumulated weight of hoop test evidence is unlikely to be strong, except at key steps in a causal chain. So, distinguishing the degree to which evidence is something you absolutely need (hoop test) from something you want (smoking gun test) but do not necessarily expect to find, helps you to understand whether you can adequately reject alternative explanations and confirm the main explanation you are trying to test. To put it in Fairfeld and Charman's (2017: 11) words, not failing hoop tests whispers in favour of a hypothesis but passing a smoking gun test allows you to shout in favour of a given hypothesis.[4]

To illustrate, let's use a simple example relevant to international development. If one wanted to demonstrate that a meeting between a specific group of stakeholders took place, an attendance sheet for a meeting would likely exist. These are not always taken, so they are not necessary, strictly speaking. However, because they are extremely common, one would generally expect to find them. Finding such attendance sheet would be an example of passing a hoop test. On the other hand, one might not expect to find photos of events as this may not be a common practice, but photos are typically unique (i.e. closely linked to the event) and thus can add some confidence regarding both the fact that a meeting happened with key

[3] The most explicit effort in methodological research is Humphreys and Jacobs' (2015) Bayesian Integration of Quantitative and Qualitative data (BIQQ) model. This will soon be tested for evaluation in the Centre for Excellence for Development of Impact (CEDIL) programme.

[4] Beach and Pedersen (2019: 42) argue that formulating rival claims is unnecessary and relies on counterfactual reasoning rather than making inferences based on processes that link together causes and outcomes. For significant outcomes such as war, there is rarely just one cause. Most outcomes in social science have more than a single cause. Nonetheless, we suggest that the exercise is useful in terms of defining degrees of uniqueness which help increase causal leverage. Establishing rival claims need not be a matter of necessity and sufficiency from a single actor or even group of actors. A claim that one step in a causal chain is necessary to an outcome is enough to suggest that an intervention had significant merit. This does not mean that other contextual and causal factors were unimportant. In this, the task is to include actions from other actors that may be part of a causal package (complementary claims), and where deemed relevant establish rival claims as a potential to refute one's own hypothesis.

individuals in attendance and potentially show that specific actions took place. Finding such photos might be an example of a passing a smoking gun test.

Contribution Tracing offers three key advantages: the capacity to assess the strength of evidence; a tool to quantify confidence; and guidance for the evaluator on what evidence to seek out, based on its probative value. The method relies on a more explicit use of Bayes theorem, which is presented in Box 2 below.

Box 2. Bayes theorem in Contribution Tracing

Bayes theorem is used to calculate our posterior confidence in a contribution claim based on our prior confidence - set at 0.5, the Bayesian "no information" tradition – and a likelihood function which relates to the difference between the true positives rate (sensitivity) and the false positives rate (type I error).The sensitivity of an item of evidence relates to the probability of observing it, if the contribution claim is true. Hoop test evidence is an example of evidence with high sensitivity. Our expectation of observing hoop test evidence is high, assuming the contribution claim is true. Therefore, not observing such evidence, lowers our confidence in a claim.

The type I error of an item of evidence relates to the probability of observing it, if the contribution claim is NOT true. The higher the type I error (value closer to 1), the less unique that item of evidence is in relation to the claim under investigation. We focus on identifying evidence with low type I error (value closer to 0). This is akin to smoking gun evidence in Process Tracing, as evidence with low type I error is unique to the claim under investigation. The larger the difference between sensitivity and type I error for an item of evidence, the higher its probative value.

Stedman-Bryce *et al*., 2017

In the following section, we will present the six evaluation cases with their core similarities and differences in how they applied Process Tracing.

iv) The Evaluation Cases

The selected cases illustrate a number of different forms and adaptations of Process Tracing. Three of the evaluations employed Contribution Tracing, and three used Process Tracing in combination with other methods (Contribution Analysis, Outcome Harvesting, and Contribution Rubrics). While the initiatives varied widely in terms of the sectors covered – from infrastructure, value chains, education, to climate change – all of them focused on governance or advocacy. The evaluations were for four organizations (and their partners): CARE International, World Vision Canada (WVC), Oxfam America (OUS), and the International Institute for Environment and Development (IIED). The evaluations took between 3 and 12 months, depending, chiefly, on the clarity of ToCs and availability of project teams to gather and process data.

The evaluations also varied in terms of the level of participation of evaluators and evaluation users, from fully external, to external participatory, to partner-led (in this context, this means the evaluation is led by implementing organizations of the initiative under evaluation). In this latter category, implementing teams took a lead role in managing and coordinating data collation, analysis and reporting, however, the process was facilitated by an external evaluator, supported by a semi-external quality assurer to help increase rigour, validity and quality (see Pasanen *et al.* 2018 on partner-led evaluation and see summary in Figure 3 below).

Each case involved the development, or refinement, of a ToC in a workshop; developing contribution claims and mechanisms; and gathering data to investigate the validity of these

claims. The processes generally involved a participatory evaluation workshop, a documentary review, Key informant interviews (KIIs), and/or Focus Group Discussions (FGDs).

Key similarities and differences between the evaluation approaches are represented below in Figure 2. In the following sub-sections, we provide further information on each of the cases.

Figure 2. Variation in Evaluation Approach

Initiative/ Agency	Method	Participation	Sector/Theme	Level
GSAM (CARE)	Contribution Tracing (mechanisms as systems)	Partner-led	Infrastructure & social accountability	Local
JATRA (CARE)	Contribution Tracing (mechanisms as systems)	Partner-led	Infrastructure & social accountability	Local
Conference of Parties (COP) Advocacy (IIED)	Adapted Contribution Tracing (without Bayesian updating)	External participatory	Climate change policy & advocacy	Global
Cocoa Life (CARE)	Contribution Rubrics (mechanisms as systems)	Partner-led	Value chains & participatory planning	Local
G7 Policy Advocacy (Coalition of WVC, Plan, Save the Children Canada, Right to Play, Results Canada, UNICEF)	Process Tracing (mechanisms as systems)	Partner-led	Girls' education policy & advocacy	Global
G7/G20/COP Advocacy (OUS)	Process Tracing (minimalist mechanisms)	External	Climate change policy & advocacy	Global

a. Ghana's Strengthening Accountability Mechanisms

Ghana's Strengthening Accountability Mechanisms Project (GSAM) was a five-year (2014-2019) USAID-funded initiative whose overall goal was to strengthen citizen oversight of capital development projects to improve local government transparency, accountability, and performance. GSAM sought to improve capacities for planning and accountability of local government officials, civil society organizations (CSOs), and citizens in 100 districts across the country. GSAM was implemented by a consortium that included CARE International in Ghana, IBIS in Ghana and the Integrated Social Development Centre (ISODEC). In 50 districts GSAM introduced public audits with the Ghana Audit Service and in a further 50 districts GSAM introduced social audits. The latter was the interest of the evaluation. For the social audit, GSAM members supported local-level CSOs to regularly monitor and gather information on the planning and construction of selected capital development projects led by District Assemblies. This included the strengthening of CSO and citizen capacities for using social accountability tools (e.g. community scorecards), the monitoring of capital projects, and the dissemination of this information to the public.

The main contribution claim that the team chose to evaluate was the following:

> *Citizen oversight over capital projects, as a result of the GSAM project activities, has improved accountability of local government authorities in the delivery of their capital projects.*

The mechanism had a total of 12 components across three causal chains[5]:

1. CSO oversight through report cards (causal);
2. District assembly engagement (causal);
3. Citizen voice through scorecards (causal).

The project team interacted with citizens in four communities situated in two of Ghana's administrative districts – Afigya Kwabre and Wassa East – that were selected as the study areas out of the 100 districts in which the project was implemented. The team also interacted with local government officials, identified as key informants, as well as staff of CSOs that were implementing the GSAM project in these districts.

Data was collected by the project team and analysis of results was conducted by a member of CARE International UK (Tom Aston), with assistance from Pamoja Evaluation Services (Gavin Stedman-Bryce, Alix Wadeson and Bernardo Monzani).

b. Journeys to Advancing Transparency Responsiveness and Accountability

The Journeys to Advancing Transparency Responsiveness and Accountability (JATRA) was a three-year (2014-2017), World Bank-funded project which took place in 15 Union Parishads (UPs) in three *Upazilas* of two districts in northwest Bangladesh. UPs are the lowest tier of local government in Bangladesh, while *Upzilas* are administrative units similar to counties in the United Kingdom. JATRA aimed to strengthen the UPs' public finance management systems to be more transparent and aligned with Bangladesh's Local Government Act (2009).

In order to achieve these aims, JATRA sought to:

- Build the capacities of citizens, especially the poor and marginalized, to engage in budget planning and implementation;
- Increase the access that citizens have to critical information, including through open budgeting processes led by the UP;
- Introduce key social accountability processes that strengthen citizen voice in decentralized development.

The main contribution claim that the team chose to evaluate was the following:

> *JATRA's facilitation of poor citizens' engagement has led to greater budget allocation of their demands in Union Parishad annual budgets.*

The mechanism had 23 components across six causal chains[6]:

- Presenting poor people's demands in budget meetings (causal);
- Authorities' engagement with poor people (causal);
- Authorities' self-assessment (causal);
- Authorities' budget planning and preparation (causal);
- Community scorecards (complementary);
- Social audits (complementary).

[5] For a visual of the mechanism and its components, please see: https://kumu.io/AstonCARE/tocs-and-network-maps#ghana-mechanism

[6] For a visual of the mechanism and its specific components in Ghana and Bangladesh, please see https://kumu.io/AstonCARE/tocs-and-network-maps#bangladesh-mechanism

Data was collected by the project team and analysis of results was conducted by a member of CARE International UK (Tom Aston), with assistance from Pamoja Evaluation Services (Gavin Stedman-Bryce, Alix Wadeson and Bernardo Monzani).

c. CARE International's Cocoa Life Project in Côte d'Ivoire

The Cocoa Life project in Côte d'Ivoire (phase 1) was a five-year project funded by Mondelēz International. The project was focused on cocoa value chains and chiefly on the role of community development planning. At the heart of this pillar of work was the formation of Community Development Committees (CDCOM) as citizen-led governance structures created in order to connect the voice of citizens to the lowest planning unit of the decentralized government. CDCOMs supported the development of Community Action Plans (CAPs) which defined which issues communities believe should be prioritized and also promoted collective action to mobilize resources to implement these plans. This process was designed to contribute to more inclusive governance achieved through citizen mobilization, voice, and representation. It was argued that this can make a significant contribution to the well-being of cocoa farmers and their families. Ensuring that basic infrastructure was provided and an improvement in basic service provision was considered to be crucial to sustainable livelihoods.

In this evaluation, we used "Contribution Rubrics". Contribution Rubrics is a theory-based, single case method which draws on Contribution Tracing and includes aspects of Outcome Harvesting and evaluation rubrics in order to assess outcomes and contribution to those outcomes (Aston, 2019).

The Cocoa Life team developed a ToC for the project's five pillars: 1) farming; 2) community; 3) livelihoods; 4) youth; and 5) environment. Based on this ToC, the team identified two priority outcome domains in line with the project's two main axes of work (CDCOMs and Village Savings and Loans Associations) that were considered of sufficient significance to merit evaluation, and which participants had seen materialize in various intervention areas. Only one of these (CDCOMs) is discussed in this paper.

The main contribution claim that the team chose to evaluate was the following:

> *CDCOMs influence the provision of selected essential infrastructure, enabling co-financing from cooperatives, communities, and other actors.*

Three causal chains were identified, and one complementary chain:

1. Community advocacy (causal);
2. Community resource mobilization (causal);
3. Cocoa Life convening and brokering (causal);
4. Creation of Mondelēz's "Opportunity Fund" (complementary).

The mechanism has 25 components in total, and two supporting steps from Mondelēz International through the creation of the "Opportunity Fund."

The evaluation looked at specific instances in which the outcome was believed to have materialized. As such, the evaluation looked specifically at the construction of a health centre in the *terroir* of Sikaboutou, department of Daloa, and the construction of a water pump in the *terroir* of Gozon, department of Duékoué. Data was collected by the project team, data processing was done by CARE International UK, and analysis of results was conducted by an external evaluator (Tom Aston).

d. IIED's support to the Least Developed Countries Group

The International Institute for Environment and Development (IIED) has been supporting the Least Developed Countries (LDC) Group of Negotiators in the context of the UN Framework Convention on Climate Change (UNFCCC) since 2001, with the aim of advancing the priorities and amplify the voices of LDCs in global climate agreements. The Group's 47 member states are among the most vulnerable to climate change but have limited capacity to influence global negotiations. To support them, IIED has thus been providing technical, logistical and financial support to increase the Group's engagement in formal UNFCCC sessions, to strengthen the role of key Group representatives, and to promote LDC positions in the media. Starting in 2011, IIED launched a new strategic effort to help the LDC Group to influence negotiations that were intended to lead to a new global treaty. This is what eventually became known as the Paris Agreement[7], the culmination of four years of intense negotiations where member states, coalitions and non-state actors jostled for influence in an attempt to ensure that their respective priorities would be featured in the treaty.

The approach chosen for the evaluation of IIED's efforts was an adaptation of Contribution Tracing, in combination with Contribution Analysis (see Befani and Mayne, 2014; Befani and Stedman-Bryce, 2016). However, the evaluation did not include Bayesian updating from Contribution Tracing. Conclusions were instead presented in exclusively qualitative terms.

The main contribution claim chosen for the evaluation was the following:

> *IIED's support to the LDC Group — and the LDC chair in particular — led to greater endorsement and formal acceptance of LDC positions and priorities in the Paris Agreement.*

A review of the initiative's ToC revealed that IIED was primarily pursuing its aim through three strategies (LDC Group activation and engagement; support to the Group's Chair, and media visibility). The claim cut through all of them and included six causal chains:

1. Greater coordination of the LDC Group, spurred by IIED support, led to increased and better participation of the LDC Group;
2. IIED support to the LDC Group Chair increased his capabilities to play a more prominent role in the negotiations;
3. Funding and logistical support from IIED led to increased engagement by LDC Group members in key meetings within and outside the UNFCCC;
4. The increased profile of the Chair brought greater political endorsement for LDC common positions and priorities (both internal to the group and external);
5. Joint IIED-LDC Group efforts led to greater clarity of LDCs' positions and priorities and outreach to other negotiators;
6. IIED contributed to increased media presence of LDCs positions and priorities.

Five alternative (rival) claims were chosen for the evaluation to test as well, each of which could have provided competing explanations for why the observed change took place:

1. The pressure to make a deal forced developed countries to prioritize global cooperation and move closer to LDC positions;
2. Other negotiating groups pushed for the same issues as the LDC Group and had more influence in ensuring that these were endorsed and accepted in the final agreement;

[7] https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

3. France, who played a key role throughout the negotiations and hosted the summit in which the Paris Agreement was signed, was committed to getting a solution that accommodated the positions of all blocs, including the LDC Group;
4. Support for the LDC Group from other institutions, such as the UNFCCC, led to greater endorsement and formal acceptance of LDC positions and priorities in the final Paris Agreement;
5. China was committed to getting an agreement that accommodated the positions of developing nations, including the LDC Group.

Data was collected, processed, and written up by an external evaluator (Bernardo Monzani).

e. Government of Canada's commitment to girls' education in crisis contexts

In the approximately 18 months leading up to Canada hosting the G7 in Charlevoix, Quebec, WVC, Plan International Canada (Plan), Right to Play, Save the Children Canada (Save), Results Canada and UNICEF (collectively, the G7CSO Coalition) collaborated to advocate for girls' education in crisis as a key agenda under the theme of Women's Equality and Empowerment, at the June 2018 G7. The G7CSO Coalition targeted its efforts towards the Government of Canada (GoC) specifically, to secure political and financial commitments for this agenda. On 9th June 2018, the G7 countries signed on to the Charlevoix Declaration on quality education for girls, adolescent girls and women in developing countries[8] along with 3.8 billion in financial commitments (including World Bank funding). The GoC made a significant commitment to this, as expressed through the G7 Charlevoix Declaration and its financial commitment of $400 million (the observed outcome).

The primary objective of the evaluation was to assess the available evidence and level of confidence in the contribution claim regarding the G7CSO Coalition's influence on the observed outcome, with focus on the GoC's specific commitment and leadership to it.

The main contribution claim that the team chose to evaluate was the following:

> *The G7CSO Coalition's policy influencing efforts through coalition building, direct government engagement and youth and public advocacy, secured the GOC's commitment to girls' education in crisis contexts, as identified through the G7 Charlevoix Declaration on Quality Education and GoC's financial pledge of $400 million.*

This claim reflects the logic and strategic pathways represented in the ToC, developed collectively as part of the evaluation process. From this, two causal chains were identified, and one complementary chain:

1. Insider advocacy (causal);
2. Coalition building and leadership (causal);
3. Youth and public engagement (complementary).

Due to the scope and resources for the evaluation, the team chose to undergo the Process Tracing on the first causal chain, which was agreed as the most significant one in terms of influence and contribution on the part of the G7CSO Coalition. The mechanism for the insider advocacy chain had 5 main components which were broken down into 24 sub-components for evidence testing.

[8] https://www.international.gc.ca/world-monde/international_relations-relations_internationales/g7/documents/2018-06-09-quality-education-qualite.aspx?lang=eng

The team also identified the following three alternative claims that could have provided competing explanations for why the observed change took place:

1. The Malala Fund through the G7 Gender Equality Advisory Council secured the GoC's commitment to reaching the observed outcome;
2. Other G7 Countries (not Canada) led the way to set the agenda to prioritize the inclusion of girls' education in crisis in the G7 agenda, shaped the contents of Charlevoix Declaration and pushed for increased financing of it;
3. The One Campaign, who led another coalition of Canadian CSOs under a theme of Women's Economic Empowerment, also influenced the contents in the Charlevoix Declaration on Quality Education for Girls, Adolescent Girls and Women in Developing Countries.

A participatory workshop was held to develop the contribution claim and identify 'expect-to-see and want-to-see' evidence. Data was collected and processed for the main causal chain and the three alternative claims, after which a report was co-written by an external evaluator (Alix Wadeson) and the internal Design Monitoring & Evaluation (DM&E) Manager at WVC.

f. Oxfam America and Climate Change and Energy Advocacy

Starting in 2016, OUS embarked on concerted efforts with other US-based NGOs to prevent the US from backsliding on international climate policies, in particular its support to the Paris Agreement, after the election of President Trump. These efforts were coordinated in particular through an informal group of NGO leaders, known as the Kitchen Cabinet, which under OUS' stewardship of its the Climate Change and Energy Advocacy Team (CCAT), had a prominent role in sharing information and developing common strategies. One of the main strategies developed during this period was to rally international support for the Paris Agreement, especially by other world leaders, who could offer a counterweight to President Trump in case he decided to pull the US from the accord. It was under this strategy that OUS decided to focus specifically on key international events in 2017: the G7 summit in Taormina, Italy (May), the G20 summit in Hamburg, Germany (July), and COP 23 in Bonn, Germany (November). OUS' contributions in relation to these took place both through the Kitchen Cabinet, as well as through the Oxfam Confederation (OI).

The main contribution claim that the team chose to evaluate was the following:

> *As linked to its US influencing strategy, in the lead-up to the G7 and G20 summits, OUS (via the Climate Change and Energy Advocacy Team) played a leadership role, both within OI and in broader civil society networks, on the strategy and actions that successfully influenced governments to uphold their commitment to the Paris Agreement (and associated actions) in the face of US backsliding.*

This claim reflects the logic and strategic pathways represented in the ToC for the overall CCAT portfolio, developed collectively as part of the evaluation process. It is important to note that for this case study, the evaluation was commissioned at the portfolio level with the objective to assess overall effectiveness of CCAT's work across a wide range of different thematic and technical areas. However, the evaluation team integrated Process Tracing as a way to add richness and rigour to the evaluation by selecting a case study that linked different interventions within the overall portfolio together and choosing an observed outcome to test in-depth. Process Tracing can therefore a be used as a standalone method for an evaluation or as part of a broader evaluation approach to highlight a specific case study and observed outcome, within a larger body of work being evaluated through other methods.

A total of 5 causal chains were then identified as being necessary to explain OUS' contribution to the outcome, while no alternative claims were identified or tested. The mechanism included the following chains:

1. OUS played a leading role in relevant NGO networks and advocacy bodies to develop common strategies to respond and counter US withdrawal from Paris Agreement;
2. OUS succeeded in making sure that the defense of the Paris Agreement was prioritized within OI and shaped OI's response;
3. OUS mobilized insiders (policymakers) to influence the Trump administration's decision-making process regarding the potential withdrawal from the Paris Agreement;
4. OI (and/ or OI-influenced) key messages reached G7 and G20 leaders, in particular those from France, Italy, Germany, and the UK;
5. OUS, OI and/ or OI-influenced public outreach efforts which helped to create public pressure, also through media visibility, on G7 and G20 leaders to counter US backsliding.

These causal chains were conceived as minimalist mechanisms. Evidence was thus collected to make causal process observations rather than broken down as components, as in a systems approach. However, we have broken them down into components in Figure 4 to aide comparison. Data was collected, processed and written-up by two external evaluators (Bernardo Monzani and Alix Wadeson).

v) What Have We Learned?

Learning and reflections on the application of Process Tracing based on the combined experiences drawn from the six case studies above are organized under four meta-themes discussed below. These are followed by a set of practical tips for applying Process Tracing in the context of international development programming and evaluation.

a. Participation

Collaboration makes a big difference

As developing appropriately specific causal mechanisms and assessing the probative value of evidence depends heavily on strong knowledge of the context, the involvement of stakeholders engaged closely in implementation can be very important for the successful application of Process Tracing. In our experience, close collaboration between internal stakeholders and external independent evaluators is well-suited to Process Tracing, and perhaps more so than for most other evaluation methods. While Process Tracing can certainly be applied independently by evaluators or social science researchers, we argue that for the purposes of international development interventions of the "hard to measure" type, participation is important for to this methodology to be effective. The table below (figure 3) provides useful distinctions for the level of participation in evaluations while the case studies presented in this paper reflect examples of the first three types of participation.

Figure 3. Levels of Participation in Evaluation

External, independent evaluation	External but Participatory	Partner-led	Self-evaluation/ internal evaluation
An evaluation conducted by organizations or people who are not part of or accountable	An evaluation typically led by an external evaluator but representatives of implementing organizations and stakeholders (may	An evaluation where the implementing partners are part of the design and take a lead role in	An evaluation carried out by those who are also responsible for the design and delivery of the project.

for those responsible for the design and implementation of the project (or initiative).	include beneficiaries) are involved in design, data collection and analysis of results. The degree of participation can vary.	managing and coordinating data collation, analysis and reporting.	

Pasanen et al., 2018: 7

It is also important to note that participation, while useful and essential, is also political. There are invariably different organizational priorities and biases brought forth, depending on the stakeholders involved and external dynamics; while this need not undermine or threaten the evaluation, it should be taken into consideration and mitigated.

Participation in evaluation design

Notably, the framework provided through a Process Tracing methodology was often appreciated as being considerably more participatory than what had been the norm in other evaluation efforts. In all cases, the staff involved from the commissioning organizations worked with the evaluators to choose and craft the language in their contribution claims. JATRA in Bangladesh and GSAM in Ghana were given complete freedom to choose the contribution claims they wished to evaluate and were engaged in every step of the process: they developed the claims; identified the evidence they needed; and set their own benchmarks for what 'success' looked like. Both teams chose to pursue ambitious claims because they wanted to demonstrate the high-level influence of their work. Similarly, in the G7CSO Coalition evaluation, the coalition members felt very confident in a high level of contribution of their work to the observed outcome. As such, they wished to ensure the language of the claim reflected this explicitly. In such situations, it is helpful for evaluators to provide coaching and assistance to ensure that the claims are of 'good enough' quality to be appropriately tested and ensure teams are aware of the trade-offs in choosing more ambitious claims. For example, more ambitious claims would require more effort in terms of planning, and data collection, which is not always obvious to project teams when initially drafting claims.

Participation in data collection and analysis

Once claims are selected, evaluators can play a critical role to help teams choose the most appropriate data collection tools, develop data collection protocols and to quality assure the data collection process. The resourcing available has implications for how participatory the data collection and analysis can be. For example, in the Ghana and Bangladesh evaluations respectively, CARE resourced an internal team to collect data, led by the project's M&E lead, in addition to the external evaluators. At any time during the process, there was at least two staff members engaged who dedicated time in the field and budget to conduct data collection and analysis. In the G7CSO Coalition evaluation, the internal DM&E lead was able to share data collection and analysis quite equally with the external evaluator, while it was necessary that the staff took an initial lead on sharing the secondary data as it largely originated from internal-only access. In the Cocoa Life Côte d'Ivoire evaluation, the data was collected by a four-person field team and processed by a two-person central office team, while the evaluator conducted the analysis. However, in other cases, both human and time resources for direct data collection and analysis were far more limited for internal staff, requiring more work on the part of the external evaluators and less participation in this regard. Therefore, for an evaluation to be participatory or partner-led, ensuring a two to four-person team (internally) is a good benchmark for data collection in a more participatory evaluation model. See Box 4 for more practical guidance for planning and resourcing.

The people in the room matter to triangulate stakeholder perspectives

Getting diverse perspectives in the room to figure out what actually happened really helps to gain clarity and keep the process honest. As such, we believe it is worth the added effort and time investment that comes with engaging more stakeholders in an evaluation process. The diversity of views will illuminate different parts of causal identification and help to identify and access important pieces of evidence. There is not one perfect mix to be prescribed here since it is dependent on the case at hand. However, in Ghana, for example, it worked well to include those closely involved in the implementation, an internal DM&E staff member, representative(s) of the partner organization(s), in addition to the external evaluation team member(s).

Process Tracing as a means to develop evaluative capacity and appreciation

Due to the high level of participation, the approach applied to Process Tracing approach also allows for capacity building opportunities for staff and partners engaged in the process. This was a deliberate goal in many of these cases (in addition to making the evaluation more cost-effective). In multi-stakeholder interventions of consortiums or coalitions with varying MEL capacity, offering an opportunity for staff of all partners to engage in the evaluation process can support increased MEL capacity overall. While evaluations are often thought of as independent processes conducted by external actors with results presented at the end, the nature of Process Tracing means that staff can be engaged in the different elements of the process such as ToC, contribution claim and causal chain development and evidence identification, evidence grading and even data collection. This can foster evaluative thinking, as well as an appreciation for the value of evaluations. For example, those involved in the G7CSO Coalition's workshop on Process Tracing and evaluation design expressed an appreciation for a robust and helpful process that could help in evidencing other advocacy work, which they had not been exposed to previously nor had they considered evaluation as especially relevant or relatable to their work before.

Accepting flexibility as a key requirement

While participation is an asset for this method and also brings the added value of capacity building, it requires flexibility of evaluators and the evaluation process. Increasing the level of participation within an evaluation can lead to making certain decisions to enhance a sense of ownership (i.e. not all decisions are made purely from a methodological perspective). For example, linked to the point above on the political aspects of evaluation participation, there are often inter-personal dynamics at play regarding the attachments to different parts of the work that influenced outcomes. Different stakeholders will bring their own strong views and it is often important to honour this with a degree of flexibility in order to protect and promote the participatory nature of the process and active engagement of diverse internal stakeholders. However, it is important to note that granting such flexibility holds more for formative evaluations than summative evaluations.

b. Theories of Change

Building blocks

As Process Tracing is a theory-based method, there is an important relationship between ToC and Process Tracing. ToCs are a *sine qua non* for an effective and rigorous Process Tracing evaluation. ToCs are an important means to articulate and test both the causal pathways and assumptions that we expect to interact and result in outcomes. A strong ToC also helps to increase our awareness of the context and landscape of actors who influence outcomes; in turn, this information can increase the validity and accuracy of contribution claims.

If ToCs are ready beforehand, it not only saves time in the evaluation but can also act as a guide to shape the whole evaluation. Without ToCs clearly articulated, the construction of mechanisms and their components takes considerably longer. In addition to this, a ToC combined with frequent and well-organized monitoring and evidence collected during implementation enables more efficient Process Tracing evaluation. However, at the same time, developing or reviewing a ToC with the participating stakeholders can act as a good entry point into the evaluation and build their capacity for evaluative thinking and practices.

In all our case studies, ToCs were not in place *a priori*, so we had to develop them as part of the evaluation process. This weakness was also found in various other evaluations which employed Process Tracing (see Stedman-Bryce, 2013, 2017). It cannot be taken for granted that the commissioning of a theory-based evaluation will necessarily have these building blocks in place and as such, due time must be allocated to building this foundation.

Degrees of testability

It is worth remembering that ToCs can be at various levels, from the most conceptual and abstract to real steps which are directly testable. Even when a team has a ToC for a program or a project, it generally needs to be made tighter and more specific to evaluate it well. For example, with the G7CSO Evaluation, the team had developed a broad ToC to depict key outcomes from the start of its advocacy and policy influencing efforts until the June 2018 G7 announcement on Girls' Education (the final observed outcome). However, the ToC was a simple visual model that acted as more of a timeline of key events. It did not identify different causal and complementary pathways of change and how these interacted. The team and evaluators used the original ToC as a model to build on and refine in order to isolate the key pathways of change with the specific actors and actions that were then used to conduct the Process Tracing exercise.

As mechanisms and causal chains are typically more granular than even a good ToC, conducting a systems form of Process Tracing means that claims are more empirically testable than those using most alternative methods. The risk of a highly testable ToC and causal chains, however, is that if "necessary" evidence is not found, it can seriously damage, or even invalidate, your claim. This may not be fatal, as sometimes one finds that a causal chain is actually just complementary, and the overall mechanism holds. However, if this evidence is not found at key points within the mechanism, or strong evidence is found for rival claims, then one can rule out one's own hypothesized explanation for change.

For the Ghana, Bangladesh, and Côte d'Ivoire evaluations, there was considerable similarities in many of the key steps, given the comparable focus on action plans and proposals from citizens to service providers and public authorities. Mechanisms and concomitant evidence (and probative value) remain highly context-specific; however, the components developed could serve as a heuristic guide for future social accountability work at CARE more generally.

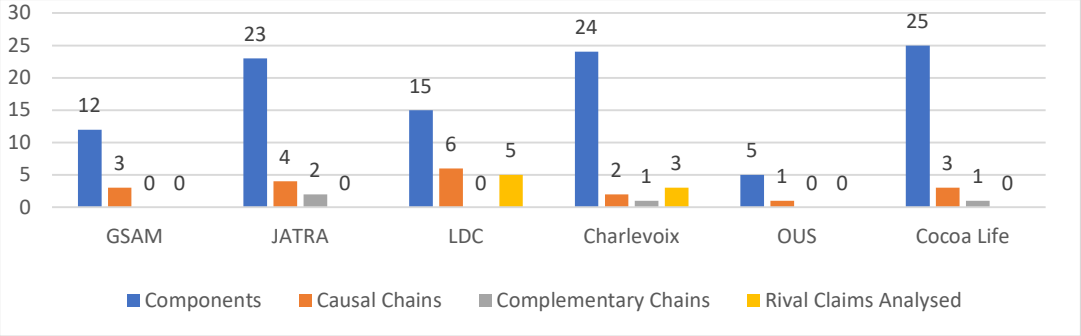
c. Methodological decisions

Causal complexity and evaluative choices

In Figure 4 on the following page, one may see that the evaluations investigated claims of varying causal complexity. What we found was that around 20 components (steps) was at the limit of what it was manageable to evaluate without risk of teams disengaging with the process. It is also at the limit of what can be achieved within the 20 to 40-day budgets most organizations have available for such evaluations.

Each component tends to entail the identification of two to five potential items of evidence. So, in practical terms, 25 components might entail the identification of between 50 and 125 items of evidence. Collecting in excess of 50 items of evidence (whether primary or secondary) is a considerable level of effort. Therefore, in the design workshop, if teams identify in excess of 20 components, it is worth considering the level of outcome and whether it may be feasible to investigate in a reasonable timeframe within a reasonable budget.

Figure 4. Causal Complexity across Evaluations



Secondly, the initial urge from teams tends to be to identify a *single* causal chain. It quickly becomes clear in most cases that there is more than one chain. Before the evaluation, the Ghana team had a process map which described, step-by-step, what they did to influence district assembly response to citizens' concerns about infrastructure investments. It was a single pathway, but throughout the process of developing a mechanism, the team realized there were, in fact, four different pathways to the same outcome (providing financing for CSOs to conduct oversight through report cards; getting district assembly members to release data; supporting citizens to monitor projects; and creating spaces for dialogue between citizens and district assemblies).

For a tacit or stakeholder-based ToC (rather than a research-based ToC as in Mahoney's [2002] discussion of critical junctures), getting diverse perspectives in the room to consider what had actually happened and for different actors to explain their reasons to others was extremely helpful in making that chain more robust and testable. The exercise also allowed the team to clearly differentiate between what was prescribed in the logical framework, what was expressed in their process map, and what they believed actually caused change (Aston, 2017). In Canada, the G7CSO Coalition identified three main chains, two of which were hypothesized to be causal and the other was judged to be only complementary. However, the group achieved consensus based on their intuitive reasoning, that only one of the causal chains (insider advocacy) was worth focusing on for the evaluation in light of time, budget, and group interests (along with the assessment of three alternative claims). This approach worked well; however, this choice and its limitations were clearly explained in the evaluation report.

However, when a team identifies more than three causal chains it is worth considering whether these are necessary or complementary. In Bangladesh, for example, the team initially identified five causal chains. In our experience, there is normally one predominant chain which comprises the bulk of effort, and this was the case in Bangladesh. However, there was disagreement regarding whether scorecards and social audits were causal (necessary) or complementary (a supportive set of conditions, but unnecessary to the claim). Strategies were iterative, but the evaluation was assessing particular years with changes in budget allocation, and thus the semi-external evaluator and external quality assurer of the evaluation (Tom Aston and Gavin Stedman-Bryce, respectively) judged that these claims were complementary. These decisions are inevitably political because they entail difficult reflection regarding the

relative weight of different strategies and their hypothesized causal power. In this example, scorecards and social audits were very important to *other* outcomes, but not necessary to the main outcome evaluated.

Thirdly, the length and number of causal and complementary chains have an important bearing on the capacity of evaluators to analyze alternative claims. Being equally tough on alternative (rival) explanations is recognized to be "best practice" in Process Tracing (Bennett and Checkel, 2014; Fairfield and Charman, 2017), and if the time available is less constrained (as is often the case in research), then this is certainly wise. Whether an evaluator is able to evaluate alternative claims is also highly political and relies to a great degree on the appetite of clients commissioning evaluations to entertain the possibility that *their* preferred explanation may be incorrect or misjudged. In our experience, clients heavily favoured the evaluation of their own claim and hypothesized causal pathways. In GSAM and JATRA, as the evaluations were chiefly formative, there was less pressure to evaluate alternatives in depth. In Cocoa Life, the request to consider the effects of a second outcome related to Village Savings and Loans Associations made the testing of rival claims unfeasible in the time available. In the G7CSO Coalition evaluation, the coalition team was sceptical about other direct influences on the observed outcome, however recognized the value in assessing alternative claims as a means to potentially strengthen their claim and rule out the explanations of rival advocacy coalitions. For the OUS evaluation, the decision to forego rival claim testing was based on resources for the evaluation rather than perceived value. And in LDC, the clear focus of the commissioning organization and their previous use and knowledge of Contribution Tracing meant the team were very open to considering rival claims in earnest.

However, the explicit assessment of type 1 error in Contribution Tracing, for example, provides a proxy for rival claims, so while it is highly beneficial to rigorously test rival claims, it may not be, strictly speaking, necessary. If a team spends considerable effort in assessing the likelihood that one may find postulated evidence even if their explanation were not true, then they can gain a good sense of whether rival claims are likely to be a threat (or not) to rule out a proposed explanation.

Evidence tests

While there is debate regarding the necessity of using formal Process Tracing tests (Fairfield and Charman, 2017), we have found that the use of hoop tests and smoking gun tests, in particular, have been very helpful to focus attention on the probative value of individual items of evidence. While some researchers and evaluators stress that Process Tracing is chiefly about testing rival claims, the majority of effort in our cases was focused on ensuring the basic credibility of proposed causal chains. The use of smoking gun tests, or simply grading evidence for type 1 error, provides a proxy to test whether rival claims may be credible in the first place. In our experience, the explicit use of evidence tests adds significant value and rigour to the evaluation. It allows evaluators to ask far deeper questions of participants about intervention context than might otherwise happen in similar evaluation approaches, given that one is explicitly asking participants to assess the chances that something else might explain the change. So, it forces participants to think very hard about "what else" might matter (see Dart, 2018 for the "what else test").

Determining which evidence will have high probative value is highly context specific. For example, in Bangladesh, one of the big questions asked of the team was about corruption. One type of corruption could be forgery of meeting records. There were meetings led by the CARE team, meetings led by citizen forums (effectively, CSOs), and meetings led by UPs (the government). Some documentation of these meetings was necessary to confirm the team's contribution claim was real (i.e. if we didn't find it, we were toast). However, the evaluators wanted to know how likely it was that we might find meeting records even if certain people had not actually attended the meetings. It turns out, due to different social norms and

incentives, the team judged that this almost never happened with CARE Bangladesh staff (due to the threat of internal sanctions). However, it often happened with citizen forums, because it was common to register people and conduct follow-up meetings after the fact with those who did not actually attend. And in government, forgery was known to happen, but it wasn't considered very common. That means that local context and varying incentives dramatically affected how the team defined the quality of evidence (Aston, 2018). In the G7CSO Coalition evaluation, the Canadian and wider G7 political context of the insider advocacy approach was highly significant to evidence tests (and grading). Decisions on the probative value of evidence such as meetings or emails required a detailed account from those team members involved in engaging with policymakers and influential political actors to explain the significance of tone, language, and access within the evidence to properly test it. For example, a direct line, even informally, to discuss the G7 policy considerations and requests for advice on the content between specific government actors and the C7CSO Coalition signalled a noteworthy level of access and credibility not readily available to others.

Grading evidence

The most important learning from two years of the Innovations in Capturing Complex Change initiative at CARE was to focus on the "right evidence." But what does this really mean? The "right" evidence is evidence that has high "probative value." In Contribution Tracing, this is assessed using Bayes formula explicitly (see Befani and Stedman-Bryce, 2017: 53 for the formula). Depending on their level of evaluation training and disciplinary background, some participants found the formula easier to understand than others. Evaluation teams did not have to learn and memorize the formula with its full notation. Instead, they had a spreadsheet with the formula and weighting of sensitivity (or certainty) and type 1 error (uniqueness) already developed by Pamoja Evaluation Services, CARE's evaluation partner. The explicit use of the formula and weighting was helpful because it allowed teams to clearly see the variation between different types of evidence and what this meant in terms of probative value. Once teams in Ghana and Bangladesh understood the exercise, they found the process very useful to defend their claims with greater confidence. However, what mattered was more the concepts of probative value and particularly that of type 1 error (uniqueness) than the explicit use of the formula in full. Further, depending on the resources available; aim of the evaluation; level of evaluative interest, and capacity of stakeholders engaged in the evaluation process, it may not be necessary, or even advisable to delve into the level of detail of presenting and using the formula. Only two of the evaluations discussed in this paper (Ghana and Bangladesh) did so.

Evidence which it would be highly unlikely to find were one's primary explanation untrue was "data gold" (Stedman-Bryce, 2017). In the Cocoa Life project in Côte d'Ivoire, as a result of the process of identifying evidence with high probative value (and low type I error), recognizing the low type I error of video evidence for the claim, they even produced new video evidence. An Extrait RTI television[9] video clip from the 10th May 2019 showed the inauguration of the health centre they had chosen as a case to test for how the project had influenced resource mobilization for infrastructure provision. Public authorities and an enormous crowd were present, and the video includes testimony affirming that the construction was due to funds from Mondelēz, the Coffee and Cocoa board and community contributions (the team's proposed explanation). So, a single item of evidence with very low type I error was a "smoking gun." In the G7CSO evaluation, an email sent to the Coalition by a high-level policy maker after the announcement of the Charlevoix Declaration thanking them for their engagement and support was ranked as doubly decisive (i.e. confirmed the claim and ruled out plausible alternatives). While in many circumstances in Canadian culture, this type of email could rationally be classified as a polite formality; however the specific tone, explicit credit provided in the language used; and the people to whom the email was directly addressed led to the

[9] See video at: https://www.youtube.com/watch?v=IrG86mTBFD8

decision to grade it at the highest level. This was an example of "data gold" or a piece of evidence one would love to find.

Gaining this grasp of probative value meant that teams could be substantially more efficient in data collection. In Bangladesh, of the 77 items of evidence identified for their causal chain, only half of them were required, because some evidence was better at validating (or refuting) the project's contribution claim than others (Aston, 2018). Conducting such an exercise early on in a project would mean that teams could also dramatically reduce the outcome monitoring data they need collect. Teams do not have to wait for the end of the project, but instead can efficiently identify and collect evidence to assess outcomes as and when they materialize, thus making evaluability easier and more efficient downstream. We found this happened in Ghana, whereby on appreciating the probative value of evidence, the GSAM team provided a cascade training to its 28 sub-contracted CSOs. In turn, they gathered data with higher probative value and particularly made use of audio-visual material, given its low type 1 error in context.

In Contribution Tracing, qualitative rubrics describe different quantitative levels of confidence ranging from no information (0.50) to practical certainty (0.99+) (Befani and Stedman-Bryce, 2016: 14). The CIA uses a scale from "certainly not, impossible" (0%) to "certainty, no question about it" (100%). These poles of certainty and impossibility are not very helpful in practice because we can very rarely be sure of anything. Indeed, if teams do not have an extremely good understanding of their ToC, it may not be possible to accurately assess levels of confidence a priori. Likewise, it is also possible for teams to game the system to get to a number of their satisfaction which accorded with a qualitative descriptor deemed acceptable. Of course, teams wanted to use words such as "certainty". The Cocoa Life team would initially use the word *irréfutable* (French, for irrefutable) as a shorthand for "data gold" in a search for evidence with low type 1 error. In reality, perhaps no claim is so solid as to be "incontrovertible", yet this conveyed the basic message for the evidence teams would dream up. Arguments happened around mid-range numbers where there was a transition between cautious (0.70 – 0.85) and high confidence (0.85 – 0.95). Using the Contribution Tracing formula, very minor variations of type 1 error had an enormous bearing on the final score. So, it might be argued that such a degree of precision is only attainable and useful if teams are sufficiently clear regarding their ToCs and think hard enough about what finding (or not finding) evidence really means in their context.

Learning from these potential shortcomings, as the Cocoa Life team also did not have a ToC at the time of the evaluation and there had been limited monitoring of outcome-level data before the evaluation, the team employed a simpler form of Contribution Rubrics with a three-point scale for hoop tests and smoking gun tests. This was sufficient to determine which evidence was most necessary and most unique. If the team had had greater clarity, they might have been able to use a five-point scale, or a ten-point scale such as that employed by the CIA between certainty and impossibility (see CIA, 1968: 5 in Beach and Pedersen, 2019: 179; Aston, 2019). The basic logic merely requires that type I error is weighted more heavily than sensitivity. In the simplest terms, evidence that can rule out your explanation should be taken a lot more seriously than evidence that can help support your explanation. As a rule of thumb, the more unique your evidence is to your intervention, the better.

Evaluative judgements and burdens of proof

This leads us to a reflection on evaluative judgements and relative burdens of proof. The choice of qualitative rubrics and thresholds for high, medium, and low confidence is also highly context specific. On one hand, different countries and cultures associate different numbers with different degrees of confidence. For example, there is a significant difference in university grades between the United Kingdom and Canada. In the UK, 70% is an A and therefore is a very good grade. In Canada, on the contrary, 70% is only a low B and therefore a relatively

mediocre score. This means there is a psychological attachment to specific numbers. And this should guide where an evaluator sets thresholds.

In Bangladesh, there were clear organizational incentives both to set an extremely ambitious contribution claim and to demonstrate a very high level of confidence. It became clear throughout the evaluation that some form of compromise was necessary. Either the team needed to lower the level of their claim (over which they would have a higher-level confidence in excess of 90%) or they should accept that higher-level claims are likely to have a lower level of confidence (around 70%). Put simply, the more ambitious one's claim, the more likely it is that other actors played a role in achieving that outcome. Thus, the uniqueness of one's contribution is increasingly threatened the more ambitious the claim.

Equally, as Process Tracing often resembles the way in which judgements are reached within the legal profession, it is worth briefly reflecting on variation in burdens of proof. In the UK, for example, there are two standards of proof in trials. The first is "beyond reasonable doubt". This means that there is effectively no reasonable doubt, or that it is simply implausible for a reasonable person to doubt. The CIA refers to 90% confidence as "beyond reasonable doubt" (CIA, 1968: 5 in Beach and Pedersen, 2019: 179). Yet, 10% is quite a lot of doubt. And in Contribution Tracing, reasonable certainty is between 0.95 – 0.99. So, there is still some room for doubt. In fact, a recent study showed that between 3 to 5% of persons convicted of capital crimes such as murder and rape were exonerated, and the figure is estimated at 6% of the criminal population overall (Berger, 2018). However, the civil standard in the UK is "the balance of probabilities" and is often referred to as "more likely than not". This maps to 0.50 – 0.70 in Contribution Tracing ("more confident than not") and to "on balance, somewhat more likely than not" (60%) for the CIA. This significant variation demonstrates that there really is no clear common standard that works across contexts. There is always potential for error, and we should only estimate levels of confidence based on thresholds agreed with participants well versed in the local context. Overall, judgements with extremely high levels of confidence (e.g. >90%) should also be viewed very cautiously.

d. Mitigating bias

Confirmation bias (type I error)

By introducing the concept of uniqueness (i.e. type I error) in relation to the probative power of evidence, evaluators can give teams a tool to isolate and minimize bias in their evidence. By working with teams to ask themselves (continuously throughout the process) the likelihood of a piece of evidence being present if the claim were not true, we can help prevent the inclination to include evidence that is biased or irrelevant to the claim and thus reduce the risk of false positives. Rather than explaining these concepts in formulaic language using terms such as type I error, a simpler way is to ask, for each component, "what do we need see if X is true?" (i.e. our minimum expectations to pass the hoop test) and "what would we like to see if X is true?" (i.e. the evidence that will convince us the most and help to rule out alternatives to pass the smoking gun test). Mapping this out within a simple document and tagging all evidence by component with degrees of "need-to-see" or "like-to-see" is a useful exercise that is accessible for most, if not all, people involved. As a brief example, if an international development organization claims to have regular access to important decision and policy makers within government as a key means to asserting its influence, one would need to see evidence of access and interaction. This could be in the form of email exchanges, meeting minutes where both parties were present, invitations to the organization to attend high-level government-hosted events, etc. If one cannot find any such evidence, the access of the organization would be reasonably called into question, and the hoop test would be failed.

Independent of evidentiary tests, the participatory nature of data collection does entail some risks of confirmation bias in data collection. Building on the Qualitative Impact Protocol (QuIP), the recently developed "Veil of Ignorance" form of Process Tracing (VoiPT)[10] offers a potential advantage in this regard for summative evaluation (Copestake *et al*., 2020). QuIP separates the role of principal evaluator/researcher from evaluation/research assistant. In this method, the principal evaluator is be responsible for developing the ToC with the project team, evaluating the probative value of evidence (*ex-ante*), interpretation and write up of findings. The evaluation/research assistant is blindfolded; they do not have knowledge of the theories, hypotheses and mechanisms being tested, only the outcomes of interest. They are responsible for selecting secondary sources, evidence collection and coding. This partition wall between principal evaluator and data collector would help reduce potential confirmation bias inherent in theory-based evaluation. It would also help allow deductive and inductive forms of Process Tracing to meet in the middle.

Evidence selection

The type of evidence one choses can help to mitigate confirmation bias. For example, open sources such as public statements contain more positive bias. Therefore, confidential sources are generally better (Beach and Pedersen, 2019: 211). Testimonial evidence in Process Tracing is best analyzed by including the source in the definition of the evidence (Fairfeild and Charman, 2017). For example, an article in a left-leaning newspaper reported "X", which thus implies certain motivations and if, then, because statements. Whether we expect to find such evidence depends upon who said it and what we believe their incentives may have been to say it or not. As another example, government acknowledgement of direct civil society influence on policy decisions is less expected, especially in certain contexts where civil society government relationships are tense. Therefore, an email, testimony, or even a public statement, which we would not expect, that attributes strong credit to civil society would hold more weight than a source in which a government actor refutes the influence of civil society (in a context of poor relations), as we would expect this and it may not be the most credible evidence to select. It is also important to note that access to evidence may be challenging in some cases, depending on what evidence is required and valuable in a given context, especially to mitigate bias. Therefore, it is prudent to consider this carefully when selecting an observed outcome and contribution claim, in order to assess feasibility. Involving more stakeholders can also help to increase the likelihood of accessing the right evidence.

Email evidence

Emails are particularly useful as they provide a timestamped interaction that identifies key actors and are generally difficult to falsify. Analysis of the tone and language of an email can provide strong potential insight into the quality and nature of a relationship between stakeholders while a series of email threads between different stakeholders can provide a reliable paper trail in the chain of events within a decision-making process. For example, work amongst advocacy coalitions often takes place on email to work on drafts of talking points or briefs, negotiating language or priorities. However, the outputs of the final documents from these multi-member coalitions do not provide the level of detail in terms of who led, influenced them, and the level of engagement or the specific challenges in coming to the final product. It is through analysis of the "behind-the-scenes" emails that we can shed light on these nuances and capture the contextual details.

This type of evidence is also relatively low in terms of cost and time; however, it depends on internal stakeholders' comfort with sharing such documentation and understanding which emails to filter for these purposes. Hence, gaining access and trust are other added values of including enough relevant stakeholders in the evaluative process as participants.

[10] https://researchportal.bath.ac.uk/en/publications/the-veil-of-ignorance-process-tracing-voipt-methodology

Testimonies from key informants

Testimonial evidence is crucial for Process Tracing, especially if there are gaps in documentary or open source evidence available. However, it is important to remember that for testimonies, the emphasis should be on quality not quantity; furthermore, emphasis on good interview skills is critical. This can require capacity building if internal stakeholders are taking part in data collection. Also important is ensuring an appropriate balance in the range of key informants selected and assessing their specific motivations vis-à-vis the given claim. For instance, selecting actors who have incentives to validate the contribution claim holds less probative value those who are considered more neutral (such as bellwethers) and those who do not benefit from the validation of a contribution claim. We should trust sources that go against the motives we would expect. For example, in the G7CSO Coalition evaluation, other civil society actors were working on different advocacy aims in the lead up to the 2018 G7, instead of girls' education. The civil society atmosphere was quite contentious at that time due to competing interests for a G7 international development commitment. The testimonies civil society members external to the Coalition confirmed the strong influence that the G7CSO Coalition members had on the observed outcome. These testimonies had a higher probative value than the members of the Coalition themselves, who already believed strongly in their influence and had incentives to do so.

Rival/Alternative claims

One of the great benefits of Process Tracing evidence tests is the possibility to investigate and potentially rule out rival claims. However, in practice, this is not always possible. For IIED, one of the authors interrogated as many as five rival claims while the G7CSO Coalition processed evidence to test the contribution claim against three rival claims. Not only can it support bias mitigation and strengthen one's claim (should rival claims be invalidated), it is also helpful for external perceptions of rigour and balance in a given evaluation. By demonstrating careful consideration of other influences for a claim, trust and transparency in one's own claim and findings can be strengthened. However, it is often more challenging to collect the breadth of evidence for rival claims than one's own claim. For example, access to other (rival) program's internal documents, communications or essential key informants may not be possible. Therefore, analysis of rival claims can result in heavier reliance on secondary or publicly available evidence (which can bring certain bias). The use of credible and relevant bellwethers is also a good option for key informants as one means for assessing rival claims.

Participation and bias

When teams are able to define their own claim, they are better able than (most) evaluators to explain how change likely happened. Project teams can definitively be involved as part of an action research process, without dramatically damaging credibility. However, they need critical friends. Therefore, incorporating measures for quality assurance, external controls or "blindfolding" are important. In addition to the QuIP VoiPT approach explained above, which offers a robust blindfolding option, there are several different ways to achieve this such as outcome panels; peer reviews; including people from different teams within the organization to participate; and a combination of an internal and external evaluation team to conduct the evaluation (as referenced in Figure 3 above from Pasanen *et al.* 2018). While the six cases presented in the paper all demonstrated degrees of participatory processes with the engagement of internal stakeholders, they also all included evaluators with varying roles of involvement. This decision was both for their expertise in the proposed methodology as well as their role to mitigate bias and ensure a balanced perspective.

vi. Recommendations to improve practice and use

The following recommendations are a compiled based on our learning and experience to date. They offer a snapshot of our current thinking, but we hope this will evolve in dialogue with other evaluators who have employed variants of Process Tracing.

R1: Context, context, context: As reflected throughout this paper, the context of a given initiative and its stakeholders is vital to Process Tracing across all facets, from stakeholder involvement to evidence selection, to evidence grading and determining probative value. While it is possible to create heuristics based on similar types of initiatives (see R5 below) to avoid 'reinventing the wheel', making all evaluation decisions through a highly context-specific lens is essential to the success of this method and the quality of analysis. This also links to the importance of a strong ToC as a foundation for Process Tracing evaluation. A context-aware ToC should identify the causal pathways, assumptions, actor dynamics, and interrelationships at play in each unique context.

R2: Highly participatory Process Tracing is worth the effort, with controls for bias integrated: Linked to R1, Process Tracing is a methodology that can benefit greatly from stakeholder participation - different staff working across an implementing organization and partner representatives too, if applicable. While this can create extra time and effort, on the part of organizations and evaluation teams, it brings benefits to the evaluation process and its outcome. Pursuing a participatory form of process tracing provides the benefits of building evaluation capacity, ownership and utilization. The key stakeholders to involve depends on the intervention in question, but it is a good idea to include a mix of perspectives and capacities with inputs from more technical staff and from those who manage and implement the programming directly. Mitigating bias remains important and can be supported by including assessment of rival claims; ensuring enough diverse stakeholders are involved; using evaluators as critical friends; peer review of the evaluation report by those not involved in the process; blindfolding for data collection; and being transparent about any methodological concessions made in the evaluation.

R3: Evidence tests and rubrics to achieve practical rigour: The quantification of confidence through Contribution Tracing offers some benefits because it can help elicit a more granular explanation. Where teams have a strong ToC (*a priori*) and sufficient evaluative capacity, this step is a valuable addition which can help increase causal leverage. However, formal evidence grading through Contribution Tracing is not always feasible for evaluators or stakeholders due to issues of time and accessibility (conceptually, language barriers, etc.). We argue that a decent level of rigour can still be achieved by using the Process Tracing tests for evidence grading and a set of simplified rubrics to assess confidence levels. Such an approach is likely to be more practical for teams with lower technical M&E capacity and can enhance utility of both the evaluative process and findings.

R4: Integrate elements of different complementary approaches to enhance evaluation practice and quality: We see significant potential for blending different methods with Process Tracing. In particular, borrowing from Realist Evaluation (Pawson and Tilly, 1997), we see advantages of making the "reasoning" of actors more explicit through *because* statements to underpin actor and activity descriptions in mechanisms. Likewise, we see value in some degree of blindfolding from VoiPT (see Copestake *et al*., 2020). While a participatory process is extremely helpful for developing causal chains and for formative evaluation, for summative evaluation, it is not necessary for project teams themselves to gather data. Resources permitting, if data is collected by local researchers with partial or full blindfolding, this can help further decrease the potential for confirmation bias. Indeed, we have also found that outcome statement templates from Outcome Harvesting can help make contribution claims more

specific and thus more testable. There are various other methods that can likely enhance the value of Process Tracing and vice versa. What matters is finding the right fit.

R5: Document and share experiences using Process Tracing: We can only stand to gain by promoting transparency and dialogue on evaluation findings and processes, including the challenges we face in conducting evaluations. While evaluation publication is becoming more common within the international development sphere, it is still relatively minimal. As there are especially limited publicly available evaluations applying Process Tracing, sharing and discussing our work openly in different forums amongst evaluation practitioners can support improvement and use of Process Tracing (in all forms) across the sector. We should also continue to identify ways in which Process Tracing can be used to support higher-quality program implementation, monitoring and adaptation, in addition to evaluation. This paper and the selected case studies did not examine the use of Process Tracing in full as a tool to support better monitoring or adaptive management, for example, by creating monitoring systems linked to ToC causal pathways thereby collecting evidence throughout an intervention that can be used to test contribution claims on outcomes, using Process Tracing Tests. We have also not explored how a Process Tracing evaluation can help to inform the design of a ToC for the next phase of a project (or a different, but similar project). These are all areas of which we encourage more exploration and documentation, as we and others, continue to apply Process Tracing to evaluate 'how' and 'why' change happens.

Top 10 tips for effective application of Process Tracing

Bennett and Checkel (2015: 21) proposed 10 "best practice" steps for Process Tracing research.[11] These are all useful cues, and several have guided us. Going beyond these lessons, we recommend the following 10 practical tips for Process Tracing evaluation to help both evaluation practitioners and implementing teams to make the most of the method.

1) **Up front time investment:** Taking the time to develop a clear ToC and causal mechanisms with project teams is crucial and should be adequately accounted for in the evaluation work plan and resourcing. However, doing so can save you a considerable amount of time in data collection downstream.

2) **Clearly define concepts:** Loose definitions for key concepts like "responsiveness" and "accountability" must be unpacked in order to be assessed. There are also translation issues with certain metaphors (smoking guns, straws-in-the-wind). So, it is important to consider language for effective communication within a given evaluation context.

3) **Clearly identify key stakeholders:** As with any evaluation process, it is important to map out key stakeholders, but in Process Tracing this is also important to help identify and rule out potential rival claims and the influence of other stakeholders.

4) **Define reasonable boundaries of influence:** You cannot evaluate everything. It is important to make choices which ensure the evaluation of the number of components (steps), causal and complementary chains is feasible with the available resources.

[11] 1) Cast the net widely for alternative explanations; 2) be equally tough on alternative explanations; 3) consider the potential bias of evidentiary sources; 4) take into account whether the case is most or least likely to alternative explanations; 5) make a justifiable decision on when to start; 6) be relentless in gathering diverse and relevant evidence, but make a justifiable decision when to stop; 7) combine process tracing with case comparisons when useful to the research goal and feasible; 8) be open to inductive insights; 9) use deduction to ask: "if my explanation is true, what will be the specific process leading to the outcome?"; 10) remember that conclusive process tracing is good, but not all good process tracing is conclusive

5) **Grade your evidence early:** Grading evidence for expected key outcomes in project design can save significant time, energy and money required for collecting monitoring data. This should also provide an excellent platform when conducting an evaluation.

6) **Gather only what you need:** More is not always better. You should only collect data appraised as having high probative value, linked to your ToC and causal pathways.

7) **Develop interview skills:** As interview evidence is often crucial, it is important that evaluation teams develop these skills. You may need to probe further to achieve sufficient depth of explanation, but you also need to limit potential biases to ensure claims are credible, particularly if your evaluation is participatory.

8) **Map your interview evidence:** Before interviewing, map out the specific components that are relevant to each key informant. This will help to develop and prioritize questions that are the most critical for component validation and Process Tracing tests. It helps interviewers make the most of limited available time with informants. It can also be helpful for a peer or external person to conduct some of the interviews for those stakeholders that are potentially too close to the team undertaking the work.

9) **Tag evidence to time and location:** It is important to have a clear chronology of the process, particularly for advocacy evaluations. If there is more than a single example of the outcome, ensure to show the location to corroborate outcome materialization.

10) **Be explicit about the why:** Motivation is key to explaining behavior change. You should thus take care to ask interviewees to explain their rationale for choices (through open questions) rather than simply assuming these are linked to the intervention.

Box 3. Practicalities of applying Process Tracing

Stakeholder buy-in and communication: The level of stakeholder involvement that Process Tracing often requires of a commissioning organization is important to clarify up front to manage expectations. As evaluators, it is critical to ensure buy-in by senior staff and those with whom you will work directly due to the level of effort involved. While this does not have to mean high-cost evaluation budgets or endless workshops, it likely means a little more staff time involved than the average evaluation as much key information on context and evidence is often best known by implementers.

Human Resources: This depends on the level of participation desired and the staff time and budget available. In our experience, a team of 2-4 people (including the primary evaluator) should be working throughout the evaluation to provide information and support for the different steps.

Capacities: It is important to include internal staff and partner representatives with different capacities. It is not essential for all participants to be experienced in M&E; it is more important that they are well-versed with the intervention; understand the context and actors involved; bring thematic or technical expertise related to the intervention; and can offer different perspectives as a way to mitigate biases and offer critical insight on potential evidence and its probative value.

Budgeting: It is difficult to put a price tag on Process Tracing evaluation as it depends on the key outcome(s) chosen, outsourcing of external expert evaluator time, and the type of data collection required. In a scenario with a dedicated internal staff member to lead with 1-2 other staff members to support it, and the use of an external evaluator (for about 25-30 days), a Process Tracing evaluation can be done for between $15,000-25,000 USD. This includes budget for a bit of travel but does not include staff time costs.

Timeframe: This is inherently linked to the human resources involved, their availability to work on the evaluation, and the complexity of outcome(s) evaluated. In our experience, a Process Tracing evaluation usually takes between 3 to 6 months with team members who can dedicate a few days each month to the process. It is important to note that the front-end time to develop the ToC, causal mechanisms, and identify the right evidence can be more time intensive than the later stages.

vii. Bibliography

Aston, T. (2017). "Howe to Avoid Toolsplaining: Thinking Differently about Social Accountability," CARE International, available at: https://insights.careinternational.org.uk/development-blog/how-to-avoid-toolsplaining-thinking-differently-about-social-accountability

_____ (2018). "Sounding Clever or Being Smart? How to do more with less in evaluating governance programmes," CARE International, available at: https://insights.careinternational.org.uk/development-blog/sounding-clever-or-being-smart-how-to-do-more-with-less-in-evaluating-governance-programmes

_____ (2019). "Contribution Rubrics."

Bates M. and Glennerster, R. (2017). "The Generalizability Puzzle," Stanford Social Innovation Review, available at: https://ssir.org/articles/entry/the_generalizability_puzzle

Beach, D. (2016). "It's All about Mechanisms – What Process Tracing Case Studies should be Tracing," *New Political Economy*, Vol. 21, No. 5, pp. 463-472.

Beach, D. and Pedersen, R.B. (2013). *Process-Tracing Methods: Foundations and Guidelines*, Ann Arbor MI: University of Michigan Press.

_____ (2019). *Process-Tracing Methods: Foundations and Guidelines, Second Edition*. Ann Arbor MI: University of Michigan Press.

Befani B, D'Errico S, Booker F and Giuliani A. (2016). *Clearing the Fog: New Tools for Improving the Credibility of Impact Claims*, London: International Institute for Environment and Development.

_____. and Mayne, J. (2014). "Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation," *IDS Bulletin, Special Issue*. Vol. 45, Issue 6, pp. 17–36.

_____. Stedman-Bryce, G. (2016). "Process Tracing and Bayesian Updating for Impact evaluation," *Evaluation*, Vol 23, No. 1, pp. 1 – 19.

Bennett, A. and George, A.L. (2005). "Process-Tracing and Historical Explanation," in A. Bennett and A.L. George (Eds.), *Case Studies and Theory Development in the Social Sciences*, Cambridge MA: MIT Press.

Bennett, A. (2010). "Process Tracing and Causal Inference," in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Brady, H.E. and Collier, D. (Eds.), Lanham, MD: Rowman & Littlefield.

Bennet, A. (2014). "Disciplining Our Conjectures: Systematizing Process Tracing with Bayesian Analysis," in Bennett, A. and Checkel, J.T. (Eds.), *Process Tracing. From Metaphor to Analytic Tool*, Cambridge: Cambridge University Press.

Berger, M. (2018). "Wrongful Convictions Reported for 6 Percent of Crimes," available at: https://penntoday.upenn.edu/news/first-estimate-wrongful-convictions-general-prison-population

Buffardi, A. Pasanen, T. and Hearn S (2017). "Understanding How to Measure Hard to Measure Stuff: Dimensions, Measurement Challenges and Responses, Measuring Development in Turbulent Times," 29 November 2017, Bucharest, available at: http://starea-natiunii.ro/images/conferinta_internationala/prezentari/50_Anne_Buffardi_hard_to_measure_-_final.pdf

Collier, D. (2011). "Understanding Process Tracing," *PS: Political Science and Politics*, Vol. 44, No. 4, pp. 23 – 30.

Collier, David, Henry E Brady, and Jason Seawright. (2010). “Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology," in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Brady, H.E. and Collier, D. (Eds.), Lanham, MD: Rowman & Littlefield.

Copestake, J. Gary Goertz and Stephan Haggard (2020). The Veil of Ignorance Process Tracing (VoiPT) Methodology: Version 8, Prepared for a Symposium in Qualitative & Multi-Method Research.

Dart, J. (2018). *The what else tool: A basic way to strengthen your impact claims and avoid having egg on your face!,* available at: https://www.clearhorizon.com.au/all-blog-posts/the-what-else-tool-a-basic-way-to-strengthen-your-impact-claims-and-avoid-having-egg-on-your-face.aspx

Gerring, J. (2007). *Case Study Research*, Cambridge: Cambridge University Press.

Gugerty M. and Karlan, D. (2018). *The Goldilocks Challenge: Right-Fit Evidence for the Social Sector*, Oxford University Press, USA.

Hay, C. (2016) “Process Tracing: A Laudable Aim or a High-tariff Methodology?,” *New Political Economy*, 21:5, pp. 500-504.

Humphreys, M. and Jacobs, A. (2015). “Mixing Methods: A Bayesian Approach,” *American Political Science Review*, Vol. 109, No. 4, pp.653-73.

Fairfield, T. and Charman, A. (2017). Explicit Bayesian analysis for process tracing: guidelines, opportunities, and caveats. *Political Analysis*, Vol. 25, No. 3, pp. 363-380.

_____ (2018). “The Bayesian Foundations of Iterative Research in Qualitative Social Science: A Dialogue with the Data,” *Perspectives on Politics*.

Faletti, T. and Lynch, J. (2009). “Context and Causal Mechanisms in Political Analysis,” *Comparative Political Studies*, Vol. 42, No. 9, pp. 1143-1166.

Humphreys, Macartan, and Alan Jacobs. (2015). “Mixing Methods: A Bayesian Approach.” American Political Science Review, Vol. 109, No. 4, pp.653-73.

King, F, Keohane, R. O. and Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press: Princeton.

Mahoney, J. (2002). *The Legacies of Liberalism: Path Dependence and Political Regimes in Central America*, The Johns Hopkins University Press, U.S.A.

_____ (2012). “The Logic of Process Tracing Tests in the Social Sciences,” *Sociological Methods and Research*, Vol. 41, No. 4, pp. 570–97.

_____ (2016). “Mechanisms, Bayesianism, and Process Tracing,” *New Political Economy*, Vol. 21, No. 5, pp. 493-499.

Monzani, B. (2018). *IIED support to the Least Developed Countries Group. Influencing global climate change negotiations.* IIED, London, available at: https://pubs.iied.org/pdfs/17466IIED.pdf

Naeve, K, Fischer-Mackey, J, Puri, J, Bhatia, R and Yegbemey, R. (2017). *Evaluating Advocacy: An Exploration of Evidence and Tools to Understand What Works and Why*. 3ie Working Paper 29. New Delhi: International Initiative for Impact Evaluation (3ie).

Overseas Development Institute (2018). *Measuring the Hard to Measure in Development*, available at: https://www.odi.org/events/4527-measuring-hard-measure-development

Oxfam GB. (2011). *Process Tracing: Draft Protocol*, Oxfam GB.

Pasanen, T., Raetz, S., Young, J., and Dart, J. (2018). *Partner-led Evaluation for Policy Research Programmes: A Thought Piece on the KNOWFOR Programme*

Evaluation. Working Paper 527. Overseas Development Institute (ODI), available at: https://www.odi.org/sites/odi.org.uk/files/resource-documents/11969.pdf

Pasanen, T. and Barnett, I. (2019). *Monitoring and Evaluation Tools and Approaches to Support Adaptive Management*, Overseas Development Institute (ODI).

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

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

Punton, M. and Welle, K. (2015). *Straws-in-the-wind, Hoops and Smoking Guns: What can Process Tracing Offer to Impact Evaluation?* CDI Practice Paper.

Sayer, A. (2000). *Realism and Social Science*. London: Sage.

Schmitt J and Beach D. (2015). "The Contribution of Process Tracing to Theory-based Evaluations of Complex Aid Instruments," *Evaluation*, Vol. 21, No.4, pp. 429–47

Stedman-Bryce, G. (2013). *Health for All: Towards Free Universal Health Care in Ghana, End of Campaign Evaluation Report*, Oxford: Oxfam GB.

_____ (2017). *Women's Empowerment in South Africa: Evaluation of the Raising Her Voice Project*, Oxford: Oxfam GB.

Stedman-Bryce, G., Budge-Reid, H., Monzani, B. Wadeson, A. and Creech, F. (2017). *Capturing the Effects of Inclusive Governance: Inception Report*, Pamoja Evaluation Services.

_____ (2018). *Innovations in Capturing Complex Change: Learning Report*. Pamoja Evaluation Services.

Stern, E. Stame, N, Mayne, J, Forss, K, Davies, R, and Befani, B. (2012). *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.

Vogel, I. (2012). *Review of the use of 'Theory of Change' in International Development: Review Report*, Department for International Development.

White, Howard, and Daniel Philips. (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.