Process-Tracing Methods in Program Evaluation

Derek Beach
Estelle Raimondo

IEG
Independent Evaluation Group
WORLD BANK GROUP
World Bank • IFC • MIGA

IEG Methods and Evaluation Capacity Development Working Paper Series

1818 H Street NW
Washington, DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

ATTRIBUTION
Please cite the report as: Beach, Derek, and Estelle Raimondo. 2025. *Process-Tracing Methods in Program Evaluation*. IEG Methods and Evaluation Capacity Development Working Paper Series. Independent Evaluation Group. World Bank.

MANAGING EDITOR
Diana M. Stanescu

EDITING AND PRODUCTION
Amanda O'Brien

GRAPHIC DESIGN
Rafaela Sarinho

Process-Tracing Methods in Program Evaluation

Derek Beach and Estelle Raimondo

Independent Evaluation Group

April 2025

CONTENTS

AUTHORS

Derek Beach[1]

Estelle Raimondo[2]

Corresponding Author

Derek Beach, derek@ps.au.dk

Author Affiliation

1. Aarhus University, Denmark

2. Independent Evaluation Group, World Bank Group

ABSTRACT

Process tracing is a theory-based method uniquely suited to assessing an intervention's contribution to outcomes, especially for interventions that are difficult to quantify, such as knowledge work or institution building. Because it relies on iteratively developing and empirically testing granular theories of change with a focus on processes, process tracing enhances the ability of evaluators to establish strong causal links between interventions and outcomes. Furthermore, process tracing emphasizes the explicit connections between actors, their actions, and resulting behavioral changes, offering two key advantages to evaluators. First, it provides a transparent framework for presenting and evaluating the strength of the evidence gathered. Second, it enables evaluators to derive practical lessons more easily.

The robustness of process-tracing findings, however, critically depends on how well theory and empirical observables come together. This paper explores the potential of process tracing in evaluation, providing a step-by-step guide for its implementation and discussing its advantages and limitations through an examination of a recent application of the method in an Independent Evaluation Group evaluation. The case examines the impact of the World Bank's knowledge and policy work in a middle-income country. Although the World Bank's impact comes as much from data, analytics, and advisory services as it does from the financing it provides, impacts from the former tend to be understudied and underevaluated, in part because of the lack of awareness of, and use of, process tracing and other methods.

ABBREVIATIONS

IEG	Independent Evaluation Group
pToC	process theory of change
ToC	theory of change

INTRODUCTION: THE USES OF PROCESS TRACING

Process tracing is a theory-based method for evaluating what contributed to program or policy outcomes in the cases studied. Process-tracing methods are particularly useful for evaluating how intangible interventions, such as knowledge work and institution building, have contributed to outcomes. Process tracing can be deployed either as a stand-alone method for evaluating a policy or program or as part of a broader evaluation in which it is used to examine in detail the processes that produced particular contributions to a policy or program's outcomes.

As an evaluation method, process tracing has three components: (i) a "process" theory, which is a granular causal theory that links an intervention and contribution through a process theory of change (pToC) that models, theoretically, the sequence of interactions between a particular policy or program intervention and a particular contribution to a policy or program outcome; (ii) "tracing" of the pToC by assessing empirical observables produced by the activities and links in the process; and (iii) drawing of more general lessons from the cases studied by identifying why the intervention worked in those cases and what contextual factors need to be present for it to work in analogous ways in other cases.

Process tracing is in the family of theory-based evaluation methods, together with contribution analysis and realist evaluation (Lemire et al. 2020). It has comparative strengths in theorizing and providing evidence for more granular theoretical processes that link policy or program interventions with contributions to policy or program outcomes (Aston et al. 2022; Schmitt and Beach 2015; Wauters and Beach 2018). Practitioners of process tracing develop theories and gather evidence in support of them through a continued dialogue between the empirical record of a case they are studying and a preliminary pToC, which they revise and update as they collect new evidence during the evaluation. Process tracing, therefore, helps evaluators understand how interventions have actually worked in real-world cases (Raimondo and Beach 2024).

Process tracing offers a language that enables evaluators to develop more granular, step-by-step theorizations regarding how particular interventions produce contributions to policy or program outcomes. Evaluators accomplish this by outlining, in a pToC, the *interactions* between the actors involved in an intervention and the actors affected by those interventions (Camacho and Beach 2023; Cartwright 2021). Working with more granular theories of change (ToCs) regarding the processes that link interventions and contributions offers

three main analytical benefits for evaluations. First, evaluators can make more *credible evidence-based claims about contributions* (Schmitt 2020). Detailing the chain of activities and links in an intervention in theoretical terms makes it easier for evaluators to develop testable hypotheses about the empirical observables that might be left if the pToC worked as theorized. Developing explicit, testable expectations about empirical observables provides focus for empirical fieldwork because the task becomes assessing whether the expected empirics were actually present in the case or not. If the expected observables are not found, the pToC should be revised to account for how it actually worked (or did not work). Second, an empirically validated pToC sheds more light on how a particular intervention produced a specific contribution, thereby providing *actionable knowledge* that can both help improve implementation of the intervention in the case studied (if ongoing) and inspire design and implementation in other programs (Schmitt 2020; Sridharan and Nakaima 2012). Third, by focusing on the concrete activities of key actors during key causal episodes, process tracing helps generate practical lessons about how interventions work and how they can work better in real-world contexts (Raimondo 2020).

The rest of this paper explains how process tracing works in practice. We provide a step-by-step guide to using process tracing through an examination of a recent application of the method in an Independent Evaluation Group (IEG) evaluation. The evaluation used process tracing as part of a broader evaluation of the World Bank Group's engagement in a middle-income country. The process tracing focused on whether the World Bank's data and diagnostic work in the client country, in a context of national sensitivity to outside intervention, had a positive impact on the country's policy coherence. Given the sensitive nature of the questions involved in the evaluation, we present an anonymized version of the process-tracing case study. Using process tracing, the evaluation examined the World Bank's convening and policy dialogue activities, including those with a think tank engaged by the client country, surrounding the production and dissemination of a major diagnostic report over four years. The evaluation also looked at the report's policy consequences several years after its publication. Although the World Bank's impact in middle-income countries comes as much from data, analytics, and advisory services as it does from the financing it provides, impacts from the former tend to be understudied and underevaluated, in part because of the lack of awareness of, and use of, process tracing and other methods. The IEG report was thus used as a proof of concept of the potential for process tracing to fill this knowledge gap.

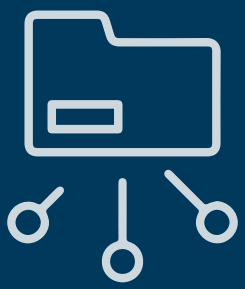
The paper proceeds as follows: Chapter 1 provides guidance on how to construct a granular pToC. Chapter 2 lays out how this pToC, once constructed, can be traced empirically. Chapter 3 identifies ways of drawing lessons from process-tracing case studies that can be applied to cases other than the one studied. Chapter 4 concludes with reflections on the applicability and limitations of process tracing.

1

THEORIZING HOW INTERVENTIONS PRODUCE CONTRIBUTIONS

Descriptive Analysis of What Happened During an Intervention

Initial Exploration of How the Contribution Could Have Been Produced



Disaggregation of the Process Theory of Change

A pToC, in effect, answers a specific and important impact evaluation question: How did intervention *x* work to produce outcome *y*? A pToC combines elements of ToCs, which describe the overall causal logic of an intervention, with theories of action, which articulate the mechanisms through which particular activities produce a specific contribution. The pToCs differ from more conventional ToCs in two key ways (Camacho and Beach 2023; Raimondo and Beach 2024). First, a ToC describes, in terms of assumptions, the causal links between the inputs to and activities of a project and the outputs from and outcomes and impacts of the project (for example, Mayne 2017, 2019). In contrast, a pToC tries to break down each of the links into a set of activities and interactions that illustrate the actual behavioral link between an input and an output. Second, ToCs tend to depict interventions in snapshot form, describing inputs and activities followed by outputs and outcomes. A pToC offers a more dynamic picture that breaks down the interactions between program actors and beneficiaries into a series of episodes.

A good pToC avoids the extremes of either drowning in detail or simplifying so much that the workings of the intervention are not transparent. An excessively detailed pToC would not be practical or useful in an evaluation because producing it would require finding and assessing evidence related to (almost) every action and interaction by every actor during the implementation period of a program or policy. An additional reason is that it would make learning across cases difficult because the pToC would include so many details that the theory would in effect be unique to the studied case. In contrast, an oversimplified pToC would not provide the theoretical scaffolding required to make credible evidence-based claims about an intervention's contributions to the outcomes of a policy or program; the oversimplification would obscure the activities and links that can be expected to leave observable traces that evaluators can assess empirically.

A good pToC includes only the *key episodes of interaction* that link the actions of program actors and those affected by these actions in a causal sequence that produces one or more contributions to program or policy outcomes; these episodes in a causal process are also known as *causal hotspots* in the literature (Apgar and Ton 2021). To identify key episodes, evaluators must ask themselves which interactions have been critical in enabling actors to *overcome challenges and barriers* that otherwise would have stood in the way of achieving the change the program or policy intended to accomplish. The pToC can also depict these challenges or barriers, together with corresponding episodes whose series of actions could, in theory, have overcome them.

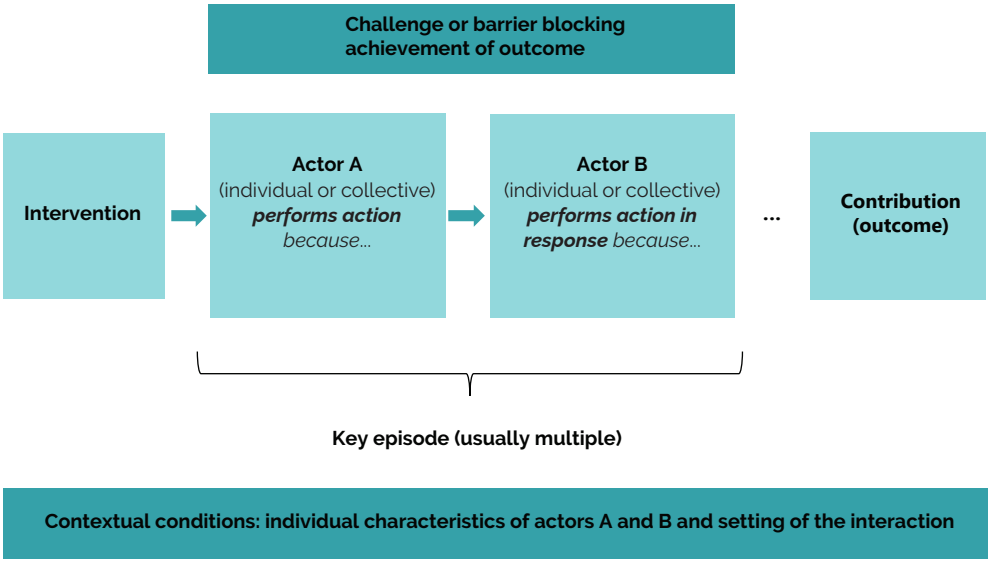
When developing a pToC, it is important to not include everything that occurred during the implementation of the program or policy being evaluated. As process tracing tends to be resource intensive, a pToC should focus on only one or a small number of key episodes, even for the most complex interventions. Evaluators will

need to consider how feasible it is to evaluate a particular pToC with the resources they have available (Aston and Wadeson 2023). Focusing an evaluation on the most interesting parts of an intervention, the most important challenges and barriers, and the contribution pathways used to overcome them is a way to avoid drowning in excessive details. A pToC will never be a perfect representation of what actually happened in a particular intervention; it will always be a simplification of reality. But it should be detailed enough that it gives those who read it a better understanding of how the intervention worked in the real world.

Within each key episode in a pToC, evaluators can use the language of *actors* and *actions* to conceptualize the interactions that constitute a process of change (Schmitt and Beach 2015; Wauters and Beach 2018). In simple terms, this means identifying *who* is doing *what*. A pToC typically presents the actors in more abstract terms that depict the causal roles they play in the process rather than using their formal titles and names. In IEG reports, for instance, we use terms such as "World Bank official" or "national ministry official."

Identifying *what* actors are doing is not enough to enable evaluators to understand why the actions of one actor led other actors to do the things they did. That level of understanding requires that evaluators supplement the process tracing with what Cartwright and Hardie (2012) term *causal principles*, defined as reasons why a given action by one actor might plausibly lead another actor to do something. A pToC can be visualized in generic terms, as in figure 1.1, in which each part is composed of an actor (who), action (what), and causal principle (why), formulated using a "because…" clause, along with the contextual conditions that evaluators might expect to find if the pToC is to work as theorized.

Figure 1.1. A Generic Process Theory of Change



Source: Adapted from Camacho and Beach 2023.

There are three steps in developing a pToC for how an intervention worked (Camacho and Beach 2023):

- Descriptive analysis of what happened, identifying what activities were associated with the intervention and what potential contributions to the final policy or program outcome each activity might have produced;
- Initial exploration of how each contribution to the final outcome might have been plausibly produced, including identification of key episodes;
- Disaggregation of the pToC into episodes of interaction between actors.

Step 1: Descriptive Analysis of What Happened During an Intervention

A pToC is a theoretical model that aims to capture the dynamics of *how* a contribution was actually produced in a particular case, based on empirical research. Before theorizing *how* a specific intervention produced a particular contribution in an individual case, however, it is vital to know something about *what* happened, empirically, in the case. Therefore, the first step in generating a pToC is to describe analytically what a particular intervention was, what activities were performed, and what potential contributions to the final policy or program outcome could

have resulted. Evaluators can undertake the necessary analysis through desk research involving program documents, but ideally they will supplement this research using exploratory interviews with program actors and managers to gain a better understanding of what actually happened during implementation of the intervention.

The initial empirical research should result in a timeline of potentially important moments in the implementation of the intervention, as well as initial hypotheses about what contribution to program or policy outcomes the intervention might plausibly have produced.

In the example IEG evaluation, we attempted to assess whether policy dialogue in the client country based on World Bank data and diagnostic work made a positive contribution to the coherence of policy in that country and, if so, how this contribution was achieved.

As the first step in our evaluation, we produced a timeline of major events in the country related to reform processes. The timeline included domestic events such as unrest and protests, political events such as major speeches or statements, and the production of data and diagnostic work by the World Bank and other relevant actors, such as domestic think tanks, with a view to identifying windows of opportunity for reform and the World Bank's potential contribution to any reforms that were achieved. To reconstruct the timeline and pinpoint key events in the case we were studying, we used as our data sources desk research that

- Identified relevant factors in the client country's political economy, including slowing economic growth and patterns of domestic unrest;
- Pinpointed central events, including analysis of speeches delivered by the client country's head of state, parliamentary decisions, and media data covering important events and World Bank activities during the period under review;
- Reviewed important documents, including the World Bank diagnostic report on the intervention itself, other internal documents related to the intervention, diagnostic reports by other actors (including a domestic think tank that had reached conclusions similar to those in the World Bank's diagnostic report), and the final reform policy document produced by the country.

On the basis of this preliminary timeline, we identified, as one potential contribution to reforms in the client country, World Bank ideas included in the diagnostic report that could have helped shape changes in the country's development model. However, at this stage, it was merely a working hypothesis that these ideas were an

actual World Bank contribution to these changes, and we would subsequently subject our hypothesis to critical empirical scrutiny and further theorization.

Step 2: Initial Exploration of How the Contribution Could Have Been Produced

After figuring out what happened and identifying potential contributions to policy or program outcomes, an evaluation can start to explore how the intervention might have produced the identified contributions. This step involves moving beyond the timeline developed in step 1 to identify what might plausibly have linked particular interventions and specific contributions to outcomes. This in turn involves identifying what plausible barriers to a particular desired change might have been present in the case being studied and trying to pin down which causal pathways could have driven the interactions that overcame those barriers. The initial theorization that results is the product of both systematic searches of existing literature (academic literature and other evaluations) and initial empirical fieldwork involving a small subset of participants in the intervention.

On the basis of a review of existing academic research and evaluations of other cases involving similar issues relating to policy advice and dialogue, evaluators should try to identify plausible challenges or barriers to change in the case they are evaluating. Doing so enables evaluators to then make a more targeted search of the existing social science literature, as well as of findings of other evaluations, and to identify potential pathways that plausibly describe overall how those challenges or barriers could be overcome. Social science theories and evaluations of other cases can be very helpful in identifying such potential pathways.

For instance, a common barrier to policy influence is getting national policy makers to take notice of ideas presented by other actors. Before this particular barrier can be overcome, policy makers must recognize that they face a problem; they may then be willing to listen to ideas about how to fix the problem. We noticed early in the IEG evaluation that national discussions about the problems the client country faced focused on disparities in poverty from region to region within the country. However, we found when we examined the existing policy documents, speeches, and parliamentary debates that they did not recognize broader problems with the country's development model during this time period, despite slowing growth and increasing inequalities.

At this stage of the evaluation, we therefore searched existing social science literature, focusing on the specific barrier of failures in problem recognition. One relevant dynamic that we identified in the policy studies literature for overcoming this barrier was incorporating benchmarking techniques in a knowledge product

(de la Porte et al. 2001; Mahon and McBride 2009). As the literature discusses, an unexpectedly positive (or negative) comparative ranking in a benchmarking exercise, presented either directly to policy makers or through the media, can draw attention to a problem. We then explored whether there was evidence of benchmarking techniques having been used to get policy makers in the client country to notice World Bank ideas relevant to solving the problem. In the case we were studying, we found that the World Bank leveraged its management of global benchmarking data, specifically indicators from *The Changing Wealth of Nations* reports, to open a window of opportunity for starting a process of policy dialogue in the country about possible reforms. In this particular case, it was the positive ranking of the client country compared with the rankings of its neighbors when intangible capital was included in measures of national wealth that prompted national authorities to initiate a dialogue on how the country's development model could make better use of human and institutional capital in the country's growth and development trajectory. We concluded, therefore, that the World Bank's use of benchmarking to overcome the lack of problem recognition among the country's policy makers was the first key episode in the case (illustrated in figure 1.2). Had this barrier not been overcome, the process would have stalled, and no amount of policy dialogue based on data and ideas could have succeeded if policy makers had remained unaware of the problem or had not felt compelled by the data and the implications of those data to ask the World Bank to provide expert diagnosis.

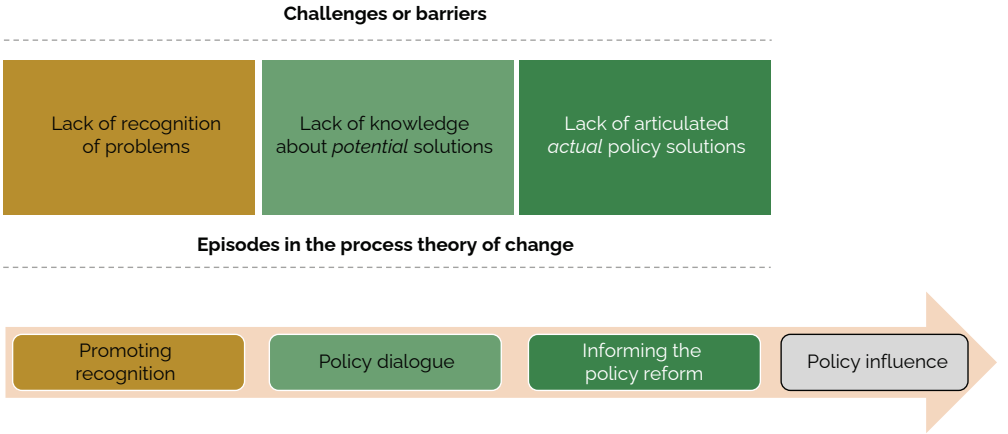
Depending on the complexity of a particular policy or program, there can be multiple challenges or barriers to change that form key episodes. In a specific case, there might, for example, be a sequence of key episodes in which one challenge is overcome, producing an intermediate outcome (contribution), followed by the next key episode, defined by a new challenge. In more complex interventions, there can also be multiple causal pathways operating in parallel.

In the IEG evaluation, to enable us to develop initial hypotheses about what types of processes might have linked data and diagnostics with policy influence, we also undertook a round of exploratory interviews with key participants to map what happened in the interactions between the World Bank and national officials during the intervention. On the basis of these interviews, we developed another initial working hypothesis focused on a later stage of the process, involving policy dialogue between World Bank officials and key national policy makers in the client country.

Through this combination of systematic literature searches and exploratory empirical fieldwork, we identified several challenges and key episodes in sequence during the time period studied. Figure 1.2 depicts these challenges and key episodes. A lack of problem awareness among policy makers can be overcome, in theory, through actions related to promoting problem recognition. Once policy makers recognize a problem,

they face the next challenge: how to fix it. In the corresponding phase of the process in the case studied here, there was still the question of how to supply the expert diagnosis of the problems and solutions. The World Bank needed to strike a balance between writing its diagnostic report in an overly diplomatic manner that would fail to help the country reform and overstepping national red lines that would make the country's policy makers dismiss the World Bank's ideas out of hand. Intense policy dialogue enabled the World Bank to overcome this challenge in supplying diagnosis that could speak hard truths without going too far (see further discussion of this later in the chapter). Finally, after it supplied the diagnosis, the World Bank still faced the challenge of getting these ideas and this advice into the country's final policy reform document without triggering national sensitivities about outside intervention.

Figure 1.2. Initial Aggregate Process Theory of Change for the Independent Evaluation Group Case Study



Source: Independent Evaluation Group.

Step 3: Disaggregation of the Process Theory of Change

Once they have identified key episodes in a case, including the challenges and barriers to change and one or more plausible pathways for overcoming them, evaluators should break down each episode into a more granular, step-by-step process. This process needs to explain *how* the specific case moves, *causally*, from the original situation to the expected intermediate or final program or policy outcomes, and what is the contribution of the activities and links involved in the interactions that overcame the barriers. Evaluators can attempt to identify through desk research and initial fieldwork what types of interactions might have taken place to overcome these barriers. Additionally, empirical examples from existing evaluation studies might provide ideas about what actions and reactions might

have taken place. As another key source, evaluators can explore relevant social science literature related to how particular challenges or barriers might be overcome (for example, how to get data noticed). Finally, evaluators can engage in a logical brainstorming about how particular challenges in the case could plausibly have been overcome, by looking either forward or backward from the initial intervention in the case to the potential contribution to final program or policy outcomes.

In practice, the formulation of a pToC often involves a back and forth between empirical observables and theory. This means that the distinction between step 3 and subsequent empirical testing can become blurred in practice. Evaluators begin with hunches about a possible pToC for a particular case—based on either logical brainstorming or case knowledge—and then engage in a preliminary round of fieldwork to probe the plausibility of these hunches. At this stage, the pToC will be incomplete, and several actions or causal principles in parts of episodes may remain unknown. Putting hunches and question marks into their preliminary pToC flags for evaluators everything that they do not yet know and that they should therefore make the focus of their empirical probing, in fieldwork, by asking, "How did it work?" Initial hunches often prove wrong, and when they do, evaluators should update their preliminary pToC to match what they have actually found in studying the case. If possible, they should engage in exploratory fieldwork to assess their initial pToC and, based on the findings of this initial fieldwork, revise and assess the pToC more systematically in a second round of fieldwork.

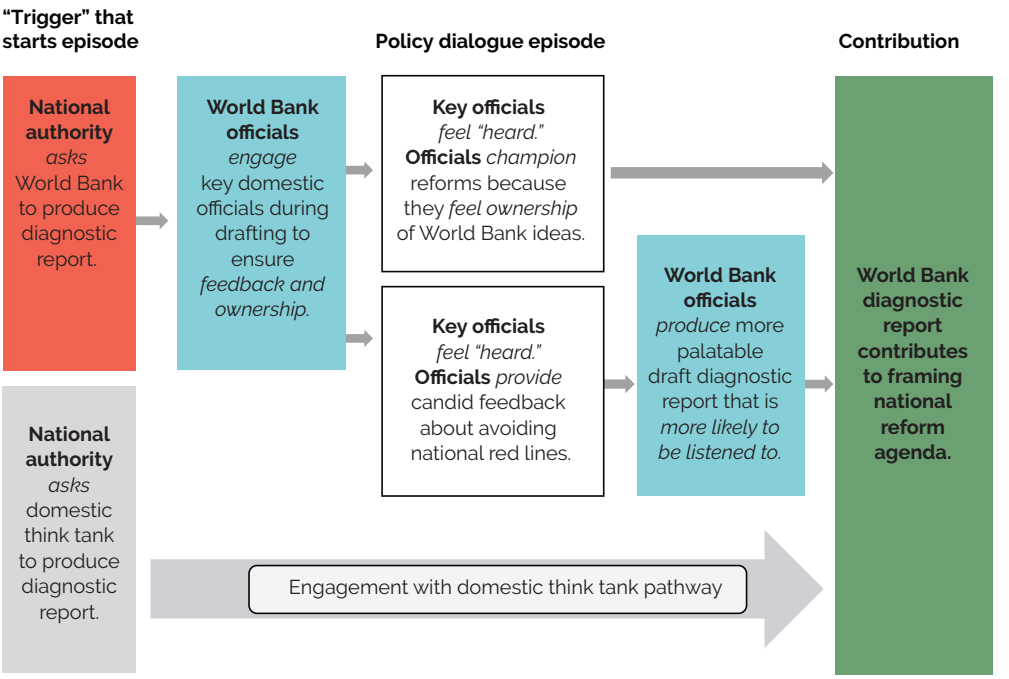
Figure 1.3 illustrates how, in the IEG evaluation, we theorized how the policy dialogue episode in which the challenge of lack of knowledge about possible solutions among national policy makers was potentially overcome. The episode involved an iterative dialogue between our initial hypotheses, the existing academic literature on policy learning and dialogue, and empirical fieldwork of the case.

In this step, we wanted to know what activities World Bank officials engaged in, with whom they interacted, and how these interactions were conducted. In the initial pToC for the policy dialogue episode, the intervention operated through two parallel pathways. We originally thought that these parallel pathways were rival contribution theories, but through empirical testing using interviews, we realized that they actually worked together in parallel, reinforcing each other, to explain the intervention's final contribution to policy reform. Here we describe the pToC for the top part of figure 1.3, which involved national authorities asking the World Bank to produce a diagnostic report of the development situation in the country. We look at the pToC for the bottom part of the figure (engagement with a domestic think tank in the client country [grayed out in figure 1.3]) when discussing operationalization in chapter 2.

As shown in the top part of the figure, we found through exploratory interviews that once World Bank officials gained a "seat at the table" to discuss with key economic

policy makers in the client country the shortcomings of the country's existing set of policies and explore alternatives, other challenges for influencing policy surfaced. These challenges included finding the right framework for focusing the policy agenda—a framework that would speak hard truths and propose sufficiently ambitious changes while not crossing red lines that would make it easy for national policy makers to dismiss the report. Further probing using additional interviews and internal project team documents revealed that engagement by World Bank officials involved first identifying potential supporters or "champions" of reforms who could make a difference within governmental deliberations, then engaging in multiple conversations with World Bank officials to sound out what areas were national priorities and to identify red lines. Through this engagement, key national officials felt "heard," leading them to (i) champion reforms along the lines identified by World Bank officials, because as the officials provided input and advice, they began to feel ownership over the ideas, and (ii) provide feedback that the World Bank team could use to revise its diagnostic report to make it more palatable to the country's decision makers (and thus more likely to be taken into account in final policy decisions). The final outcome was a reform agenda that was partly shaped by the contribution of the World Bank (the diagnostic report).

Figure 1.3. Policy Dialogue Episode (Engagement with National Officials)

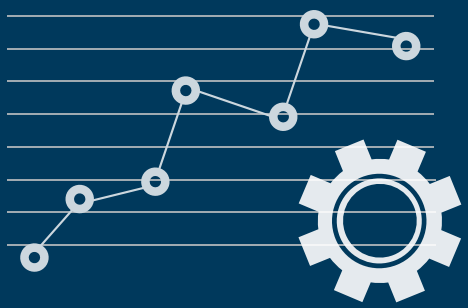


Source: Independent Evaluation Group.

2

TRACING A PROCESS THEORY OF CHANGE EMPIRICALLY

Operationalization of Expected Observables



Fieldwork to Test and Revise the Initial Process Theory of Change

Once they have developed a preliminary disaggregated pToC, evaluators can engage in further fieldwork to more systematically assess whether the pToC worked as theorized or, if it did not, how they should revise the pToC. This involves what can be termed "operationalizing" the theory, in which testable hypotheses are developed for what types of empirical observables might be left if the activities and links played out as theorized in the pToC. In simple terms, operationalizing a pToC involves asking questions such as "If activity and link A took place, what type of empirical traces might we expect that they left in the case?" The hypothesized empirical observables can also be considered the "fingerprints" that might be left by the activities and links in a case.

In process tracing, empirical evidence can be any form of empirical material that changes evaluators' confidence in how a particular theory worked in the selected case. Evidence can be *sequences* of events in a case, *patterns* in the empirical record (for example, the number of downloads of a report), *traces* in which mere existence provides proof, or *accounts* from interviews and the content of documents (Beach and Pedersen 2019). Different research techniques are relevant for collecting and assessing different types of evidence. Note that this can include statistical analysis of patterns in the empirical record, if relevant.

Process tracing often involves different modes of empirical research through the course of an evaluation. As discussed in chapter 1, the development of an initial pToC for an intervention involves an initial round of considering and probing the empirical record. A more systematic testing and revision phase then follows that involves (i) operationalizing the pToC in the form of expected observable traces that are tested empirically and (ii) assessing the collected evidence and, if necessary, revising the pToC.

Step 1: Operationalization of Expected Observables

The working pToC for a particular case being studied should be operationalized by asking what empirical observables the actions and links in the case might have left. Process-tracing methods build on Bayesian logic, in which evaluators update prior confidence in the workings of a theory based on new evidence they have gathered (Beach and Pedersen 2019; Befani 2021; Befani and Stedman-Bryce 2017). Evaluators can increase or decrease their degree of confidence in a theory based on this updating, depending on whether they have found confirmatory or disconfirmatory evidence. Bayesian logic suggests that some empirical observables can be characterized as "need to find" because the action or link for which they are expected to provide evidence should have left a particular fingerprint in a case (Befani 2021).

Not finding that fingerprint would disconfirm to some degree evaluators' confidence that the action or link is present in the case being studied.

Other empirical observables can be characterized as "love to find," meaning that if found, they provide relatively strong confirmation of the actions and links involved in the pToC for the case being studied because no plausible alternative explanations for finding the evidence exist. These observables can therefore be thought of as a confirmatory "signature" that the part of the process for which they provide confirmation is working as theorized (Befani 2021), but if other explanations for finding an observable are equally plausible, then finding the observable provides little or no confirmation of the actions and links involved in the pToC. In addition, if not found, love-to-find observables do not necessarily invalidate the pToC. These terms are defined further, alongside examples from the IEG evaluation, in box 2.1.

Box 2.1. The Confirmatory and Disconfirmatory Power of Evidence

Need-to-find evidence = disconfirmatory evidence. Need-to-find evidence is empirical observables that should be observed as a result of activities associated with a part of a process. If such empirical observables are not found in the case being studied, the lack of expected evidence disconfirms, to some degree, that that particular part of the process took place, with the degree of disconfirmation depending on how likely it was that the evidence would be found. Need-to-find evidence is related to terms such as *rate of false negatives*, *sensitivity* and *certainty* used in other research traditions.

- Example from the Independent Evaluation Group evaluation: If World Bank ideas helped shape policy reforms, we should expect at least some overlap between the final reforms and the ideas. Not finding any overlap at all would be very disconfirmatory.

Love-to-find evidence = confirmatory evidence. Love-to-find evidence is empirical observables that ideally would be observed as a result of activities associated with a part of a process and whose presence is difficult to explain in alternative ways. If such empirical observables are found in the case being studied, the evidence may confirm that that particular part of the process took place, depending on how unlikely alternative explanations for the evidence are. Love-to-find evidence is related to the terms *rate of false positives, specificity,* and *uniqueness* used in other research traditions.

- Example from the Independent Evaluation Group evaluation: Particular wording from a World Bank diagnostic work is found in a final reform document. If it can be demonstrated that no other actors put forward similar proposals, and the work was put forward before the final reform, finding such wording is highly confirmatory because it would be highly implausible that the group

(continued)

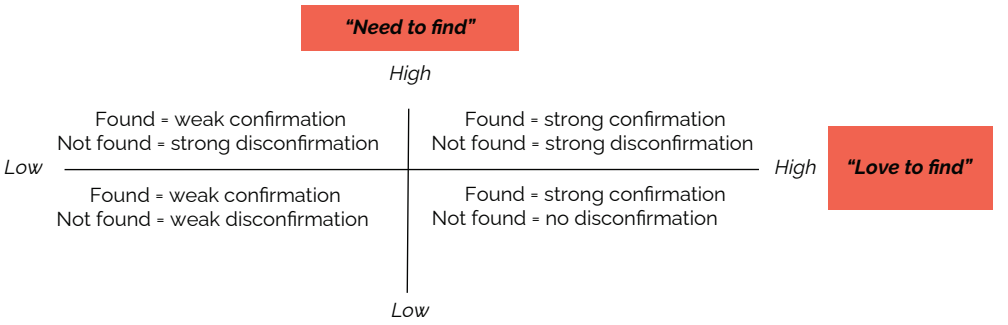
Box 2.1. The Confirmatory and Disconfirmatory Power of Evidence (cont.)

preparing the reform document would have reached such similar language purely by coincidence.

Source: Independent Evaluation Group.

Empirical observables can be both need to find and love to find, just one or the other, or neither. Evidence that is both high need to find and high love to find is confirmatory if found *and* disconfirmatory if not found. Evidence that is neither need to find nor love to find has little probative value by itself—although it might play a role in corroborating other evidence. Figure 2.1 illustrates the two dimensions and their relationship with each other. The dimensions should be understood as continuums, with some types of evidence offering stronger confirmation (if found) than others.

Figure 2.1. The Two Dimensions of Probative Value of Empirical Evidence



Source: Independent Evaluation Group.

For critical assessment of whether evidence exists that particular actions took place and were linked in the way theorized, the specifics of the pToC determine what types of sources evaluators are interested in and what questions they ask these sources. Evaluators will want to interview those people either who were directly involved in the actions depicted in the pToC or who can provide them with important (unbiased) accounts of those actions. Similarly, the questions evaluators ask in interviews should directly relate to the content of the pToC (Camacho et al. 2025). As a result, evaluators will often tailor interview questions to a particular respondent, depending on what elements of the pToC that respondent is able to shed light on.

When hypothesizing what empirical observables might be found, it is important to cast the net widely for different observables that a given activity and link might have left. Because the most probative confirmatory or disconfirmatory evidence is often not available, in most situations evaluators have to settle for second-best evidence in which each individual piece tells them little. However, if those pieces are independent of one another, when combined, they can have greater weight. Independence of evidence relates to whether sources could have influenced each other or not. For example, if we interview two colleagues in an organization and we find similar reconstructions of a set of events, it could be that these are two independent eyewitness accounts or, alternatively, that the colleagues have colluded to harmonize their answers to the evaluators' questions. Only when independence can be demonstrated can we treat the evidence as independent. Working with evidence, therefore, often involves relatively painstaking piecing together of different types of evidence (Beach and Pedersen 2019).

Returning to the IEG evaluation example, we operationalized expected observables for the action "World Bank officials engage key domestic officials during drafting to ensure feedback and ownership" by thinking about the different types of empirics that might have been left in the case by the activities involved in this action. Our expected observables involved the following:

- We expected to find, in interviews with World Bank officials, that
    - They spent time trying to identify and cultivate relevant contacts in the government of the client country;
    - They met relatively frequently with these national officials;
    - They were interested in getting candid feedback from national officials and in promoting their feelings of ownership of the issue.
- If it proved possible to access it, we expected to find, in project documentation for the production of the diagnostic report, information about meetings with national officials where ideas and potential formulations of draft text were discussed.
- We expected to find, in interviews with national officials, that they
    - Met with World Bank officials to discuss the diagnostic report, and
    - Provided feedback on the report draft.

(Despite many efforts, we were unfortunately unable to interview the national officials involved; how we managed to deal with this is discussed later in the chapter.)

We characterized most of our expected observables as need to find, although interviews with national officials might have provided us with love-to-find evidence.

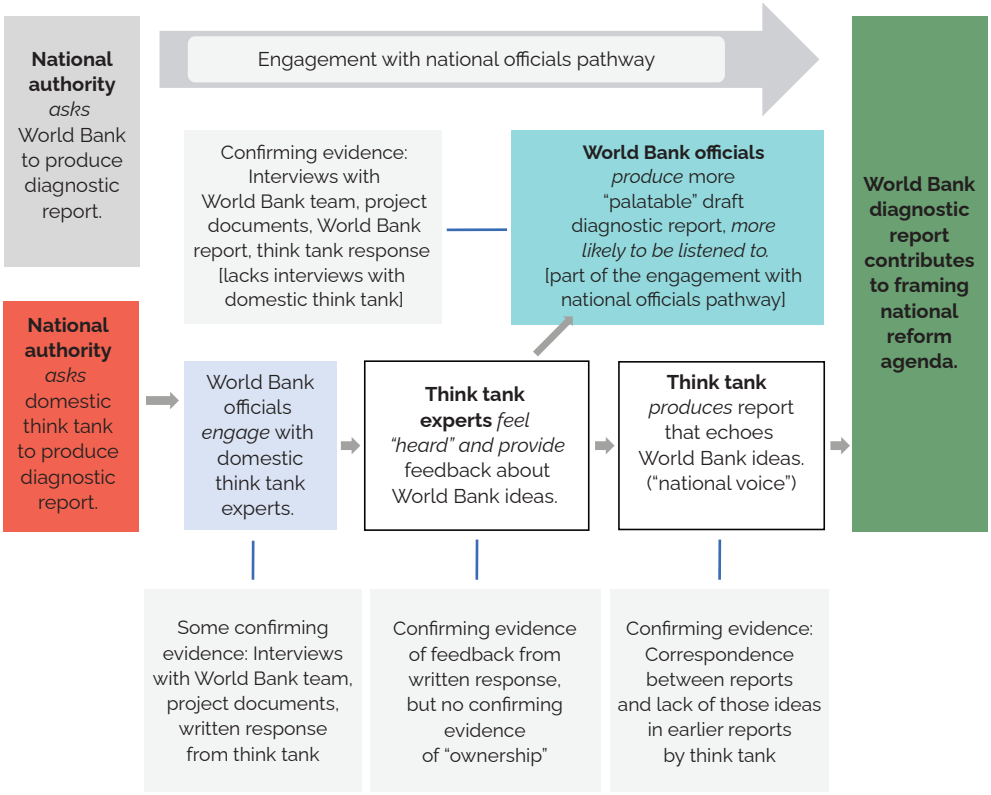
Step 2: Fieldwork to Test and Revise the Initial Process Theory of Change

In step 2, evaluators go again into the field to collect empirical material related to the expected empirical observables of the pToC. In some instances, this material does not confirm the initial pToC, with evidence suggesting that the whole process, or parts of it, worked in different ways than theorized. Such an outcome should lead evaluators to revise the pToC and thereafter produce a new set of expected empirical observables that can then be rigorously assessed through further fieldwork.

Returning to the IEG evaluation example, we broke down the second pathway of the policy dialogue episode, depicted in figure 2.2, into actors, action, and links. In the second pathway, World Bank officials were theorized to have engaged in dialogue with a domestic think tank that might have led them to adopt World Bank ideas in their own report. Engagement was theorized to involve the domestic think tank providing feedback about the suitability of World Bank ideas for potential reforms. By feeling "heard," we theorized that the domestic think tank officials might have taken some of the World Bank ideas on board and included them in their own report. In this sense, the World Bank ideas might have more credence than otherwise because they were also being promoted by a trusted national voice.

In attempting to track down evidence to assess this part of our pToC, we could not access some sources that would have been able to shed more light on how the engagement actually played out. We were able to interview the World Bank officials involved, giving us evidence of whom they met with and how frequently (once a month). Given that the information from the interviews was collected several years after the events took place, it unfortunately lacked precise details on the feedback received, as well as the specifics of the dialogue. This meant that we needed either to treat the evidence as weaker than ideal or to seek corroboration through other sources of evidence. Figure 2.2 documents the evidence for each part of the episode.

Figure 2.2. Policy Dialogue Episode (Engagement with Domestic Think Tank)



Source: Independent Evaluation Group.

In pursuing and assessing evidence for the domestic think tank side of the interactions, we sent the think tank a set of questions and received a brief written statement in response. The statement provided some evidence regarding one particular topic on which feedback was provided, but it was vague about the rest, and it was not very specific about the nature of interactions between World Bank officials and the think tank staff. Ideally, for our evaluation, we would have received more documentation regarding the domestic think tank's written and oral interactions with World Bank officials, but we were unable to gain access to such documentation despite repeated attempts. As a second-best option, our evaluation relied instead on indirect evidence of the interactions between the think tank and the World Bank (with weaker confirmatory power), where we used knowledge products produced by the domestic think tank *before* the World Bank report as a baseline for comparing the think tank's diagnosis of problems in the client country prior to engagement. We then compared early drafts of the World Bank diagnostic report with later drafts

and the report the domestic think tank produced for the client country government. We mapped when data and ideas first appeared in the reports and whether they were included or dropped in subsequent versions. In this way, we were able to establish that early think tank documents had not included many of the ideas that were found in early World Bank drafts of the diagnostic report and that were subsequently reflected in the domestic think tank's report.

If evaluators find evidence confirming a particular aspect of their pToC, they need to assess whether the evidence is enough to enable them to stop collecting data, especially if the aspect is an important causal element in the pToC. In the IEG evaluation, because our initial pToC described interactions among diverse actors, we would typically have conducted multiple interviews from different sides of the interactions (if we had gained enough access). If the accounts from different sides of the interaction had been similar and consistent, we might have concluded that we had confirmed the workings of that part of our pToC.

However, interviews in evaluations are typically with *stakeholders*, all of whom may potentially show bias because they have a stake in the project in some form or other. Therefore, evaluators should ideally also corroborate interview findings with other types of evidence (Camacho et al. 2025). If evaluators' sources are weak (that is, if we cannot necessarily trust their veracity), the evaluators should try to collect more information through other sources of evidence. Finding multiple pieces of confirming evidence helps corroborate the veracity of particular sources. In other situations where we can trust the evidence, finding one piece of confirming evidence can be enough. In the IEG evaluation, for example, we would not have expected the client country's head of state to cite the World Bank as a justification for a change in the country's policy direction. However, we did find that he had done just this, in an official speech, along with providing quite specific details about some of the problems the World Bank had identified. We took this to be a strong confirming piece of evidence that World Bank ideas were being listened to at the highest level in the country.

If evaluators do not find evidence confirming one or more aspects of their pToC, three different situations may apply: (i) they have found disconfirming evidence and should revise their pToC, (ii) they have found contradictory evidence, or (iii) they have not found any relevant evidence.

In the first case, the evidence might be very straightforward—for example, that a particular theorized action did not take place—in which case the evaluators would want to revise their pToC accordingly.

In the second case, evaluators need to figure out why the evidence they have found is contradictory. This can involve trying to collect other independent evidence that can corroborate one of the alternative interpretations of what happened and why. It is important to consider, however, that contradictory accounts do not necessarily mean one source of the accounts is wrong. It might be, instead, that the activities and links occurred in a way that was more complex than initially theorized, making both sources of evidence consistent with a revised pToC. Reconciliation of this type often requires significant detective work in piecing together the activities and links and how they worked in the case being studied. Returning to the IEG example, at the beginning of our inquiry—that is, before we identified the domestic think tank pathway—we found some contradictory evidence that suggested that the domestic think tank's ideas might have been more important in shaping the revised reform agenda than the World Bank's diagnostic report. However, with further digging, we discovered that there had been significant engagement between the think tank and World Bank officials. By tracing who put forward what ideas and when, we were able to reconcile the initial contradictory findings, uncovering evidence that through the World Bank's work with the think tank behind the scenes, World Bank officials' ideas had actually shaped the think tank's report (that is, we discovered the second policy dialogue pathway).

In the third case, not finding relevant evidence can mean different things. If evaluators cannot obtain need-to-find evidence for something after systematically searching for it, one possibility is that this absence of evidence means that the evidence does not exist at all and therefore disconfirms the corresponding part of the pToC. A second possibility is that there are important pieces of evidence that evaluators cannot access or that do not exist (for example, activities in a meeting were not recorded). Even if no evidence is available for a particular activity and link in a pToC, evaluators can expand the search by asking whether there might be indirect or circumstantial evidence for that activity and link, for example, if the inputs (actions) correspond closely with the outputs (expected responses). If even this is not possible for a particular part of the pToC, evaluators will want to establish evidence for this part even more indirectly. They can do this by finding strong evidence for the part before or after, which may shed light on the part for which evidence is lacking. Ultimately, it is still good to know what one does not know. If the lack of evidence is related to something evaluators would have loved to find rather than something they needed to find, not finding the evidence tells us little.

In the IEG evaluation, we were not able to talk directly to the national officials who advised the client country's head of state and so could not confirm that the World Bank's diagnostic report had actually made it to the advisers' ears and that they

had used it in their own advice for the head of state's speech to the nation. We did, however, have interview data from World Bank officials suggesting that they had met with these national officials. Further, we found more indirect, circumstantial evidence of the link between the World Bank diagnostic report and the head of state's speech, such as the similar framing of issues in official country documents and those of the World Bank, the fact that the World Bank framing preceded the national one, and the fact that similar frames were not found in any other published report or the like. Further, we gathered information from press coverage of interactions between the World Bank team and high-level country authorities that shed light on the extent to which World Bank ideas had framed how national officials understood the problems facing them.

Regarding the impact of World Bank ideas on the country's final policy reform document, we found relatively strong confirming evidence that indicated that a high-level authority in charge of adopting an alternative policy framework for the country's future had used the conclusions of the World Bank's diagnostic report a few years after its publication. To ascertain this, we had examined the levels of correspondence between the diagnostic report, other reports and relevant knowledge present in the country, and the final policy reform framework the country adopted. In particular, we assessed whether the other reports and the final framework showed signatures of World Bank influence in the form of particular formulations or combinations of ideas that would suggest that national officials relied heavily on the World Bank diagnostic report.

Evaluators will assess the evidence they have collected and will continue with fieldwork until they have evidence strong enough to confirm each key episode of their pToC to some degree. If significant amounts of disconfirming evidence are found, the pToC should be revised, after which the revised version should be tested systematically. Typically, a follow-up round of fieldwork is necessary in the final stages of an evaluation to fill in evidential gaps.

Bayesian logic offers an intuitive framework for assessing the degree of confidence evaluators can hold in their pToC based on the evidence they have assembled. Table 2.1 depicts how varying degrees of confidence in a theory can be expressed in words—language that intelligence agencies use widely in presenting assessments and US courts employ to summarize the strength of evidence behind their conclusions. Numerical equivalents are also presented, although we recommend that they are not used in final evaluation reports. The exception would be if the likelihoods of finding or not finding evidence can be meaningfully quantified, as some scholars suggest (for example, Befani 2021; Fairfield and Charman 2017). In real-world evaluation settings, however, *formalized* Bayesian updating is not always practical or useful.

Table 2.1. Strength of Evidence Expressed Linguistically

Strength of Evidence	Linguistic Expression	Numerical Equivalent
Strongly confirming evidence (high internal validity)	"Beyond reasonable doubt," "almost certainly"	>90% (greater than 9-in-10 chance)
	"Very probably"	80% (8-in-10 chance)
	"Probably"	70% (7-in-10 chance)
	"Somewhat more likely than not"	60% (6-in-10 chance)
	"Neutral," "as likely as not"	50% (1-in-2 chance)
	"Somewhat less than even chance"	40% (4-in-10 chance)
	"Probably not"	30% (3-in-10 chance)
(low internal validity) Strongly disconfirming evidence	"Very probably not"	20% (2-in-10 chance)
	"Almost certainly not"	10% (1-in-10 chance)

Source: Adapted from CIA 1968.

Different evaluations can require different evidential thresholds, depending on their purposes. An evaluation that seeks actionable knowledge or strong evidence-based conclusions typically requires multiple rounds of research that revise a pToC based on emerging trends in the evidence found, especially if the pToC does not operate in a way that produces empirical observables that can be collected easily. When relatively strong confirming evidence is found, the evaluation report can state that the evidence suggests that the pToC "very probably" worked as theorized in the case(s) studied. In other situations, lighter (that is, weaker) evidence supporting a relatively simple pToC might be enough for the evaluation report to conclude that it is "more likely than not" that the pToC worked as theorized. In the example IEG evaluation, we significantly updated our confidence in our theorized claim, from being rather skeptical of the World Bank's impact through a particular channel to concluding that the intervention we examined very probably worked as we had theorized based on the cluster of different types of confirming evidence that we found.

3

LEARNING LESSONS FROM PROCESS-TRACING CASE STUDIES

Case Selection

Details of the Inner Workings of the Intervention

Knowledge of Contextual Factors

A process-tracing evaluation can stand alone, with the findings confined to the bounds of the case studied. If the evaluators can find relatively strong confirmatory evidence for a disaggregated pToC, the process-tracing evaluation has relatively high internal validity. However, depending on the purpose of an evaluation, it can be relevant to ask whether the lessons learned from a process-tracing evaluation are externally valid—that is, can these lessons tell us anything about how similar contributions might be produced in other programs or policies found in similar or different contexts? One way evaluators can achieve this type of external validity is through conducting two or more process-tracing case studies of similar interventions in parallel, using the emerging pToC for each as a source of the parameters for a process-level comparison. If the evaluators find similar processes in the cases being compared, this constitutes evidence that the pToC can be applied more broadly, at least to the cases studied. Assessing similarities and differences in relevant contextual conditions between the cases can enable the evaluators to make further generalizations—albeit cautious ones—to other cases in similar contexts (Camacho and Beach 2023).

If evaluators are able to study only one case, they can still use their findings to make cautious generalizations to other cases. To determine whether an intervention that worked in one instance might also work in others, they must gather a number of pieces of information relating to various questions. Process tracing has a comparative advantage in providing answers to these questions (Woolcock 2022). For example, the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology employs a generalizability framework for its impact evaluations that seeks to explain (i) the detailed theory behind a type of intervention drawn from particular evaluation case studies, (ii) what local conditions must hold for the theory to apply, (iii) how strong the evidence is for the hypothesized behavioral change, and (iv) the evidence that the implementation process for the intervention can be carried out well (Bates and Glennerster 2017). Process tracing can help evaluators answer all these questions, but what broader lessons can be drawn from a process-tracing evaluation of an intervention depends on a number of considerations: (i) case selection, (ii) the extent to which the inner workings of the intervention are typical or unique in a given context, and (iii) the extent to which the process tracing has unpacked the contextual factors (key facts) that play out in the case studied.

Consideration 1: Case Selection

If evaluators have chosen a typical (that is, representative) case to examine, based on comparisons with other cases, then their findings provide some evidence that

the intervention studied might work in similar ways in other cases. However, an individual process-tracing case study provides evidence only of how an intervention worked within the case studied. This means that without some form of processual comparison across multiple cases, we would not know whether a selected case was "typical" or not. Another strategy can be to select cases based on a comparative assessment of the contribution involved, with evaluators selecting either "success" or "failure" cases.

We selected our case as a potential successful case based on preliminary knowledge of the extent to which World Bank work had influenced the client government's developmental policy. We also selected it because of its potential to drive lesson learning, given that it can be potentially considered a typical case of the World Bank's use of its core diagnostic work paired with policy dialogue and convening, which can be used to inform major reforms in middle-income countries. Confirming that the case was actually "typical" would require comparison with how similar interventions took place in other cases in similar contexts.

If evaluators study no additional cases using process tracing, they cannot make a firm evidence-based conclusion that the intervention they have studied will work in similar ways in other cases. For instance, although a case might look similar to other cases based on the intervention involved and its potential contribution to the results, the cases might have different inner workings, or the intervention might play out differently under different circumstances. When dealing with complex interventions whose workings depend very much on context, it can therefore be risky to *assume* that the interventions will work in similar ways in cases other than those studied (Woolcock 2022). A single case study can, however, act as a plausibility probe in relation to a type of intervention by providing evidence that it at least worked in one case (Falleti and Lynch 2009; Woolcock 2022), making it more plausible that the intervention studied might work in other cases, subject to how sensitive it is to even slight differences in context.

Consideration 2: Details of the Inner Workings of the Intervention

For all but the simplest of interventions, how things work in practice will differ to some degree in different cases. How then can an evaluation conclude that processes were similar enough in two or more cases to confirm that the intervention will work in a similar way in other cases? A good pToC should enable evaluators to make what can be termed "process comparisons" in which they use the pToC's description of actors and actions and causal principles for key episodes to assess whether the

actors and the *particular* actions they performed in a given case were *functionally equivalent* to what other actors were doing in another case.

In the example IEG evaluation, we tested (lightly), in other cases within the framework of the World Bank's country program evaluation, the lessons emerging from process tracing about the actors and actions and causal principles present in the case. We did this to check whether we could find *functionally equivalent* instances that had created similar results. We found, for example, that leveraging good data optics in the form of benchmarking and relative rankings as a very effective way of getting World Bank diagnostics noticed operated similarly in no less than four other cases. Similarly, we determined that in-depth and prolonged engagement on finding the right framing for reforms that speak hard truths while avoiding red lines operated similarly in several other cases within the same country across sectors and types of diagnostic. As a third example, we found that the use and cultivation of domestic champions of reforms and solutions proposed by the World Bank worked in similar ways in other cases within the same country.

Here, process tracing was useful in unearthing how various mechanisms played out in more detail in a case and in helping name and identify the contextual conditions under which they played out. It was then much easier to evaluate whether our pToC worked in a similar way in other instances and whether it yielded similar outcomes. Process tracing also revealed key lessons on how to draft a tailored knowledge product that questions the status quo without being so controversial that decision makers will not hear the message—for example, the need to engage both strategically and tactically with key stakeholders over a longer period than was typical when drafting knowledge products to build trust and rapport. The engagement required astute understanding of the client country's political economy, notably, a thorough reading of the stakes and power of various individuals and how to choose and cultivate relationships with champions.

Consideration 3: Knowledge of Contextual Factors

The crux of the external validity challenge in process tracing is the following: When can generalizations about how things might work be made to other cases that have not been directly studied? Even when other cases are not examined, looking at key episodes sheds light on contextual conditions that might be required for the actions and links in a particular pToC to function. Contextual conditions are factors that determine whether particular pathways within key episodes function in causally analogous ways in other cases. Making the contextual conditions and the inner workings

of interventions explicit enables readers of an evaluation to determine whether these contextual conditions are also present in other cases of interest to them.

In the example IEG evaluation, the relevant contextual conditions operated at various levels: macro, meso, and micro. At the macro level, client country authorities had a window of opportunity to rethink the development framework within which the country had operated for multiple decades because of a noticeable decline in the country's growth despite continued high levels of public investment. Lingering episodes of civil discontent that put a certain amount of pressure on the authorities to envision an alternative also contributed to opening up this opportunity. At the meso level, a long-standing trusted partnership between the country's authorities and the World Bank, as well as a tradition of "challenging each other to learn together" that had been built through decades of engagement, enabled the World Bank to have an impact through its analytical and data work. At the micro level, the inner workings that the World Bank team activated (long-term engagement on one knowledge product, cultivation of champions, and co-creation of a diagnostic framework, among others) were possible only because of a high degree of project autonomy granted by World Bank management and because of the entrepreneurial attitude of the project's team leader.

An evaluation can also probe a population of cases in a larger class (for example, all low-capacity states) more systematically using follow-up studies that trace only the most critical episodes of a pToC. Evaluators can use comparative methods, such as qualitative comparative analysis, to map key similarities and differences in contextual and other conditions across cases, enabling them to select diverse cases strategically within a population of potential cases (Raimondo 2023).

Although each of these more superficial follow-up studies may have relatively low internal validity, finding confirmatory evidence for how things work across a strategically selected set of cases can increase evaluators' confidence in the external validity of the pToC employed to study them (Beach and Pedersen 2019). If different processes are found to be operative in a particular case, evaluators should then compare it with the case previously studied to identify what differences might explain why things worked differently in the two cases.

CONCLUSION

The robustness of process-tracing findings depends on how well theory and empirical observables come together, and it can be assessed according to three criteria: (i) a disaggregated and fine-tuned pToC that captures the key episodes and mechanisms that can explain the links between the intervention studied and the outcome of interest; (ii) the finding of evidence that is unique (that is, love-to-find evidence that cannot be explained by alternative explanations) for each key episode of the pToC; and (iii) trustworthy sources of information and broad access to the empirical record. Conversely, if a pToC is too abstract or simplistic, if the evidence found to corroborate the pToC could also serve to validate an alternative theory, or if the sources employed for evaluating the pToC are biased, causal inference will be weak.

If these three conditions are met, however, process tracing can bolster evaluators' ability to provide strong evidence of causal links between interventions and outcomes, while also unveiling explanations of how and why a particular intervention triggered a specific process of change that led to the outcome of interest. The scaffolding employed for finding evidence in support of the pToC also provides a transparent way of presenting and assessing the strength of the evidence gathered and triangulating across sources. This transparency is a strength of the approach compared with other case- or theory-based methodologies. In compelling evaluators to focus on causal explanations and the links between actors, their actions, and induced behavioral changes, process tracing also makes it much easier for evaluators to derive practical lessons and ideas on how such activities can be changed to improve outcomes. Process tracing's comparative advantage over other (impact) evaluation approaches lies in its ability to assess interventions that do not lend themselves to quantification or experimentation, such as research, advocacy, and knowledge and data work, as well as policy dialogue and budget support.

Like other approaches, process tracing is not a silver bullet for solving all evaluation questions or studying all interventions. It also has some limitations to keep in mind when deciding whether to incorporate it in an evaluation design. First, it is not adequate for answering questions that require the estimation of the magnitude of a treatment effect, such as how much of an impact a particular intervention had, on average, on an outcome of interest. Second, although process-tracing principles can be intuitively incorporated into any evaluation design, the full application of the approach requires

considerable time and resources because of the need to iterate between evidence and theory, gain familiarity with how to assess the probative value of different types of evidence, and learn how to construct a pToC at a level of abstraction that is fit for the purpose at hand and how to leverage existing literature to theorize well. Third, on its own, process tracing has weak external validity and needs to be paired with a cross-case design to build in generalizability. Ideally, evaluators will trace two or more cases empirically and compare their workings at the processual level to enable them to conclude whether an intervention worked in analogous or different ways in the cases, as well as to more systematically probe the impact that contextual conditions have on how things played out in the cases studied.

When presenting the findings of a process-tracing evaluation to stakeholders, as when using the method itself, evaluators should give their pToC a central role. A simple visualization of key episodes in terms of actors, actions, and links is a good heuristic tool to help stakeholders understand how the intervention actually worked (or did not) and why. Presentations should also clearly flag the strength of evidence underlying the findings (that is, the degree of internal validity based on the strength of evidence found). The benefit of using Bayesian language to summarize the strength of evidence behind a pToC is that it clearly flags for readers how much confidence they can reasonably have in the conclusions (see box 2.1 and table 2.1).

Finally, a report on a process-tracing evaluation should clearly flag both the contextual conditions within which the evaluation's pToC can be expected to function and whether there is any evidence from other cases that the pToC works in similar ways in those cases. Without this information, readers do not know whether the findings can be applied to other cases and, if so, where the findings might provide relevant lessons for other cases. That being said, evaluation findings should be written to meet the needs of the intended users, and more often than not, this means interpreting the findings and their implications, and writing them in plain language. Methodological appendixes that clearly and transparently lay out the process-tracing approach, the pToC developed, and the evidence found in support of it can be useful in that regard.

REFERENCES

Apgar, M. and G. Ton. 2021. "Learning Through and About Contribution Analysis for Impact Evaluation." Institute of Development Studies, September 10, 2021. https://www.ids.ac.uk/opinions/learning-through-and-about-contribution-analysis-for-impact-evaluation.

Aston, T., C. Roche, M. Schaaf, and S. Cant. 2022. "Monitoring and Evaluation for Thinking and Working Politically." *Evaluation* 28 (1): 36–57. doi:10.1177/13563890211053028.

Aston, T., and A. Wadeson. 2023. "Process Tracing Innovations in Practice: Finding the Middle Path." CDI Practice Paper 25, Institute of Development Studies. https://www.ids.ac.uk/publications/process-tracing-innovations-in-practice-finding-the-middle-path.

Bates, M. A., and R. Glennerster. 2017. "The Generalizability Puzzle." *Stanford Social Innovation Review* 15 (3): 50–54. doi:10.48558/EYY5-3S89.

Beach, D., and R. B. Pedersen. 2019. *Process Tracing Methods.* University of Michigan Press.

Befani, B. 2021. *Credible Explanations of Development Outcomes: Improving Quality and Rigour with Bayesian Theory-Based Evaluation.* Expert Group for Aid Studies.

Befani, B., and G. Stedman-Bryce. 2017. "Process Tracing and Bayesian Updating for Impact Evaluation." *Evaluation* 23 (1): 42–60. doi:10.1177/1356389016654584.

Camacho, G. G., and D. Beach. 2023. "Theorizing How Interventions Work in Evaluation: Process-Tracing Methods and Theorizing Process Theories of Change." *Evaluation* 29 (4): 390–409. doi:10.1177/13563890231201876.

Camacho, G. G., J. Schmitt, and D. Beach. 2025. "Working with Interviews in Process Tracing Evaluation Methods: How to Get More of Your Interviews Using a Process Theory of Change." *Evaluation* [In press].

Cartwright, N. 2021. "Rigour Versus the Need for Evidential Diversity." *Synthese* 199: 13095–119. doi:10.1007/s11229-021-03368-1.

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

CIA (Central Intelligence Agency). 1968. *Bayes' Theorem in the Korean War*. Intelligence Report. CIA.

de la Porte, C., P. Pochet, and B. G. Room. 2001. "Social Benchmarking, Policy Making and New Governance in the EU." *Journal of European Social Policy* 11 (4): 297–307. doi:10.1177/095892870101100401.

Fairfield, T., and A. E. Charman. 2017. "Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats." *Political Analysis* 25 (3): 363–80. https://www.jstor.org/stable/26563430.

Falleti, T. G., and J. F. Lynch. 2009. "Context and Causal Mechanisms in Political Analysis." *Comparative Political Studies* 42 (9): 1143–66. doi:10.1177/0010414009331724.

Lemire, S., L. R. Peck, and A. Porowski. 2020. "The Growth of the Evaluation Tree in the Policy Analysis Forest: Recent Developments in Evaluation." *Policy Studies Journal* 48 (S1): S47–S70. doi:10.1111/psj.12387.

Mahon, R., and S. McBride. 2009. "Standardizing and Disseminating Knowledge: The Role of the OECD in Global Governance." *European Political Science Review* 1 (1): 83–101. doi:10.1017/S1755773909000058.

Mayne, J. 2017. "Theory of Change Analysis: Building Robust Theories of Change." *Canadian Journal of Program Evaluation* 32 (2): 155–73.

Mayne, J. 2019. "Revisiting Contribution Analysis." *Canadian Journal of Program Evaluation* 34 (2): 171–191.

Raimondo, E. 2020. "Getting Practical with Causal Mechanisms: The Application of Process-Tracing Under Real-World Evaluation Constraints." *New Directions for Evaluation* 2020 (167): 45–58. doi:10.1002/ev.20430.

Raimondo, E. 2023. "The Rigor of Case-Based Causal Analysis: Busting Myths Through a Demonstration." IEG Methods and Evaluation Capacity Development Working Paper Series, Independent Evaluation Group, World Bank. http://hdl.handle.net/10986/39773.

Raimondo, E., and D. Beach. 2024. "Process Tracing Methods in Evaluation." In *Research Handbook on Program Evaluation*, edited by K. E. Newcomer and S. W. Mumford, 570–86. Elgar.

Schmitt, J. 2020. "The Causal Mechanism Claim in Evaluation: Does the Prophecy Fulfill?" *New Directions for Evaluation* 2020 (167): 11–26. doi:10.1002/ev.20421.

Schmitt, J., and D. Beach. 2015. "The Contribution of Process Tracing to Theory-Based Evaluations of Complex Aid Instruments." *Evaluation* 21 (4): 429–47. doi:10.1177/135638901560773.

Sridharan, S., and A. Nakaima. 2012. "Towards an Evidence Base of Theory-Driven Evaluations: Some Questions for Proponents of Theory-Driven Evaluation." *Evaluation* 18 (3): 378–95. doi:10.1177/1356389012453289.

Wauters, B., and D. Beach. 2018. "Process Tracing and Congruence Analysis to Support Theory-Based Impact Evaluation." *Evaluation* 24 (3): 284–305. doi:10.1177/1356389018786081.

Woolcock, M. 2022. "Will It Work Here? Using Case Studies to Generate 'Key Facts' About Complex Development Programs." In *The Case for Case Studies: Methods and Applications in International Development*, edited by J. Widner, M. Woolcock, and D. O. Nieto, 87–115. Cambridge University Press.

The World Bank
1818 H Street NW
Washington, DC 20433