Article

Working with interviews in Process Tracing evaluation methods

Evaluation
1–20

DOI: 10.1177/13563890251319011
journals.sagepub.com/home/evi
Sage

Gabriela Camacho Garland
Aarhus University, Denmark

Derek Beach
Aarhus University, Denmark

Johannes Schmitt
German Institute for Development Evaluation (DEval), Germany

Abstract
In this article, we provide practical guidelines for using interviews in Process Tracing evaluation methods. Process Tracing evaluation method employs more disaggregated theories of the causal process, termed a process theory-of-change. The process theory-of-change serves as the basis for conducting and using interviews in evaluation, including in the selection of respondents and the formulation of questions. The use of interviews varies across the three different phases of data collection in Process Tracing evaluation methods: (1) an exploratory phase that develops a plausible process theory-of-change, (2) a phase for testing and revising the process theory-of-change, and (3) a phase for assessment and addressing loose ends. We present guidelines for conducting and using interviews in each phase, with a particular focus on the testing and revising phase.

Keywords
interviews, process theory of change, Process Tracing evaluation methods, theory-based evaluation

Introduction

Theory-based evaluation methods (TBEMs) put the theoretical explanation of what happened in an intervention at the center of an evaluation. The focus on developing and testing a theory

Corresponding author:
Gabriela Camacho Garland, Department of Political Science, Aarhus University, Bartholins Allé 7, Building 1331, 317, 8000 Aarhus C, Denmark.
Email: gabriela.camacho.garland@gmail.com

has far-reaching implications for conducting interviews in evaluations, both in terms of who we interview, what we ask them, and how we assess the data gathered. While there are both general guidelines for using interviews (Adams, 2015; Guion et al., 2011; USAID, 2011) and more specific guidelines for other TBEMs such as contribution analysis (CA) (Mayne, 2012) and realist evaluation (RE) (Manzano, 2016; Pawson, 1996; Pawson and Tilley, 1997), this type of guidance does not exist for using interviews in Process Tracing evaluation methods (PTEM). This reduces our ability to get the most out of interviews within PTEM.

In contrast to other TBEMs, PTEM uses more disaggregated theories of the causal process to capture the core causal dynamics of the interactions between program actors and recipients that lead to some outcome or contribution being produced (Camacho Garland and Beach, 2023; Schmitt and Beach, 2015). This type of theory can be termed a process theory-of-change (pToC), in which the process is disaggregated into a sequence of actors interacting with each other that links the intervention with an outcome (contribution). Overall, a pToC is a simplified causal explanation of how an intervention produced a contribution in a case (or set of cases), focusing on interactions between actors. Ultimately, a pToC is used to give causal meaning to empirical evidence.

Many of the implications of using a pToC as the basis for conducting interviews are similar to those for other TBEMs. These include the focus on selecting respondents based on what they can tell us about the workings of a theory, the refinement of theories during the course of an evaluation, and when to stop data collection (i.e. when there is a "sufficient amount" of confirmatory evidence in relation to the evaluation purposes (Mayne, 2012), resulting in a form of evidential saturation).

However, there are also significant differences between conducting interviews using pToCs in PTEM and how interviews are conducted using other evaluation methods. For example, a pToC tries to capture more explicitly the actions and interactions that provide the causal linkage in a process (for more, see section "Unpacking how an intervention works using a pToC"). This means there is a different theoretical focus than in CA and RE (Schmitt and Beach, 2015). In CA, a theory-of-change (ToC) describes a sequence of intervention activities, outputs, outcomes, and impacts bound together by causal link assumptions (Mayne, 2012). In RE, a ToC takes the form of Context-Mechanism-Outcome (CMO) configurations (Pawson and Tilley, 1997, 2004). In contrast, in a pToC, the focus is on actions and interactions, which has implications for who we interview and the questions we ask. Actions and linkages can be expected to leave empirical traces. Therefore, a pToC is relatively easy to operationalize in the form of expectations about potential empirical evidence that might be found in a case, both in terms of what we should find using interviews and also in other types of evidence (Camacho Garland and Beach, 2023). This means that PTEM are typically less dependent on interviews than other TBEMs. Instead, the pToC helps the analyst use other types of evidence to corroborate and supplement the data gathered through interviews.

In this article, we provide a set of practical guidelines for conducting and using interviews based on our own evaluation experiences using PTEM from a number of studies. The core idea is that working with a more disaggregated theory of the process (i.e. a pToC) helps focus the dialogue between the developing theory for how the intervention worked (or did not work as intended) and the empirics of the case. How interviews are used varies during the three different phases of data collection in PTEM. The three phases are: (1) an exploratory phase that develops a plausible pToC, (2) a testing and revising of the pToC phase, and (3) an assessment and tying up loose ends phase. We present guidelines for interviews in each of the phases—focusing in particular on the testing/revising phase.

During the exploratory phase of an evaluation, interviews with a small number of implementing actors and managers are used to help develop a plausible preliminary pToC through probing of plausible linkages. During the fieldwork phase, the preliminary pToC is central to selecting who is relevant to interview, as well as what specific and general questions we would ask different respondents. Since interviews are only one potential source of evidence, working with a disaggregated pToC helps the evaluator identify other types of relevant evidence (e.g. documents). As the goal in PTEM is to update our knowledge of how the process worked, the pToC will be under continual refinement and revision during the course of the evaluation. This means that instead of a more traditional fieldwork process in which all of the interviews are conducted and then analyzed collectively afterwards, in PTEM findings are assessed in real time in order to guide subsequent interviews in terms of potential new persons we might want to interview, or by asking other questions of respondents than what we had originally intended. In the final assessment phase, interviews can still play a role, acting as important tools for clearing up discrepancies and evidential lacunae.

This article proceeds as follows. In the next section, we briefly introduce PTEM and what pToCs look like in abstract terms as well as its main differences with RE and CA that are relevant when conducting and using interviews. The section "pToC-based interviewing—Practical guidelines" provides practical guidelines from PTEM for interviewing in the different evaluation phases. We illustrate the guidelines drawing from practical experience of applying PTEM in an evaluation of an Assisted Voluntary Return and Reintegration (AVRR) program. The conclusions briefly discuss the practical benefits of using pToCs to structure how we conduct interviews in practical, real-world evaluations.

Unpacking how an intervention works using a pToC

In this section, we briefly describe the key elements of what a pToC is and how it is used in PTEM, and how this has different consequences for interviewing when compared with other TBEMs. While PTEM is a close cousin of other TBEMs such as CA and RE, the key difference between the methods relates to the nature of what is being theorized. In a pToC as used in PTEM, the process is more disaggregated into episodes composed of actors interacting with each other that together link the intervention with a contribution (for more, see Camacho Garland and Beach, 2023; Schmitt and Beach, 2015; Wauters and Beach, 2018).

Figure 1 depicts a generic template for a pToC. The components of a pToC are the *actors* and *actions* that they perform in their interactions with each other. In Figure 1, actor A is the "sender" of an action that triggers some form of action in response from actor B. To understand the linkages between actions, we would make explicit the causal principles for *why* one action can trigger an actor to do something in response (Cartwright and Hardie, 2012: 22). Causal principles can involve intentions, motivations, and the meanings behind the actions of Actor A, and on the receiving end, how Actor B perceived the action and why it led to a particular action in response (Fujii, 2018; O'Mahoney, 2017).

A good pToC does not unpack every interaction during the implementation of a program or project. Instead, it focuses on the *key episodes*, defined as interactions between program/project actors and recipients that overcome challenges and barriers that stand in the way of the achievement of a contribution (Camacho Garland and Beach, 2023). For example, a common challenge faced by nongovernmental organizations (NGOs) advocating for policy change is getting policymakers to recognize a given policy problem. The interactions that enable an NGO to get a problem recognized would be a key episode in an advocating for policy change causal process.

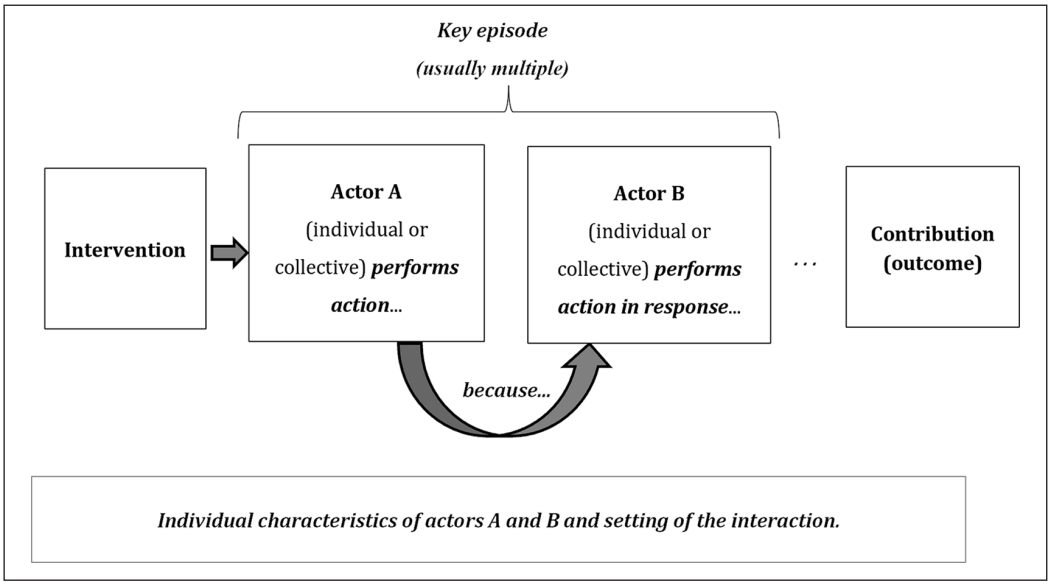

Figure 1. A generic process theory-of-change.

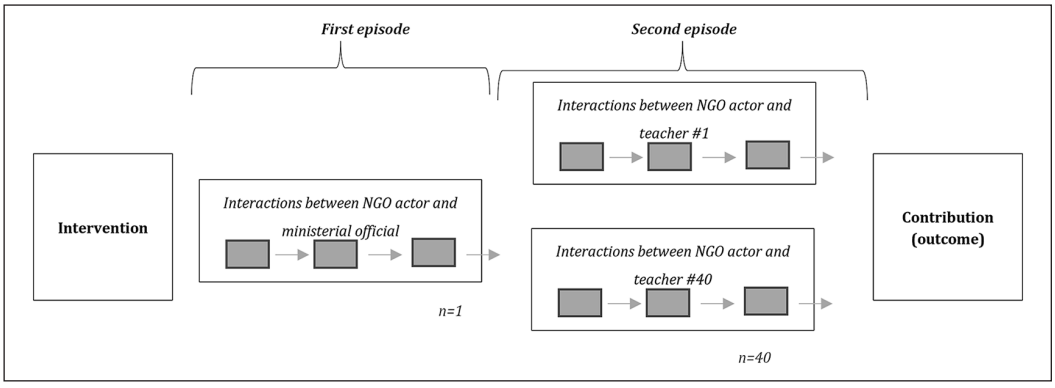

Figure 2. Differing levels of analysis within episodes of a pToC.
Source. Authors.

Within each key episode, a pToC describes the actors and actions and the causal linkages between actions that together lead to the challenge/barrier being overcome. A good pToC unpacks key episodes in enough granularity that they can be wrong empirically. However, to be tractable, they are simplified so they only focus on one or two actions/reactions. In addition, within each episode, different actors, actions, and linkages might be theorized, depending on the program/project.

Different episodes in the sequence of a pToC can take place at different levels of analysis. Figure 2 illustrates differing levels of analysis during a pToC. For instance, an early episode might deal with interactions between a project leader and an external actor (e.g. a ministerial official who grants the NGO access to a school), followed by many separate

interactions between several NGO project members and the teachers at the school. In the first episode, there would be two actors involved in the interaction ($n = 1$), whereas the second episode takes place in parallel across many instances of project member–teacher bilateral interactions, with the contribution being the product of the sum of parallel interactions (e.g. $n = 40$).

PTEM shares common characteristics with other TBEMs like RE and CA. All three methods agree that theories are refined during an evaluation, with different questions asked at the beginning than at the end. This cycle is most clearly seen in RE, in which initial CMO theories are gleaned and further developed from participant accounts, and are refined against observational and documentary data (Pawson and Tilley, 1997: 167). Later, other questions are asked during the theory refinement and consolidation phases of an evaluation (Manzano, 2016: 352). Data collection can stop when the assessment of evidence from multiple sources (including the interview data) finds that there is a sufficient amount of confirmatory evidence in relation to the evaluation purposes (Mayne, 2012).

There are however significant differences, mostly driven by how the ToCs are built and what they depict in each method. In CA, stakeholders help build a ToC, in combination with prior evidence and other sources. Interviews are then used to explore causal link assumptions (Mayne, 2012, 2015) However useful, the pathways causally linking the activities of an intervention with its intended results are not disaggregated, making it difficult to determine what evidence causal link assumptions should leave in any given case. By not depicting the specific actions and interactions, we will also have a harder time identifying precisely who is in the best position to provide us with the information needed to make a contribution claim. The questions will also differ, as the emphasis on causal packages in CA (Mayne, 2019) will inevitably lead to a different set of questions to identify them. In contrast, a pToC is relatively easy to operationalize in the form of expected empirical manifestations because the actions and interactions, as well as the actors performing them, are made explicit.

Similarly, the key differences between RE and PTEM relate to what theoretical aspects we are asking about in our interviews. In RE, the focus is on CMOs and how an intervention worked, for whom and in what contexts. Every participant can then help uncover "micro-events and processes" (Manzano, 2016: 350). Mechanisms are viewed as the causal "glue" that binds interventions and outcomes together in particular contexts. However, because mechanisms are treated as unobservable in RE, it is difficult to assess them empirically using interviews. The questions RE asks about mechanisms are typically divorced from the actions and interactions in the implementation of the program. This narrow focus on mechanisms, without clearly defining what actions mechanisms are connecting, results in evidenced theories that miss the dynamic sequential interaction between the implementers and the program recipients. PTEM offers a way to ask about intentions and perceptions more directly, thereby evaluating it instead of putting it in the assumptions. In a pToC, the process is unpacked into a sequence of interactions during key episodes. Actions and interactions can be difficult to evidence empirically. Making them explicit in our pToC enables us to operationalize them in the form of expected empirical observables, which in turn helps formulate relevant interview questions that will allow us to collect such observables.

Finally, in both CA and RE, interviews risk eliciting answers from interviewees where they speculate about mechanisms or causal link assumptions in situations where causation might not actually be there. When using a pToC, we are not assuming that unobservable mechanisms exist based on stakeholders' accounts of it, but instead we are trying to evidence the linkages

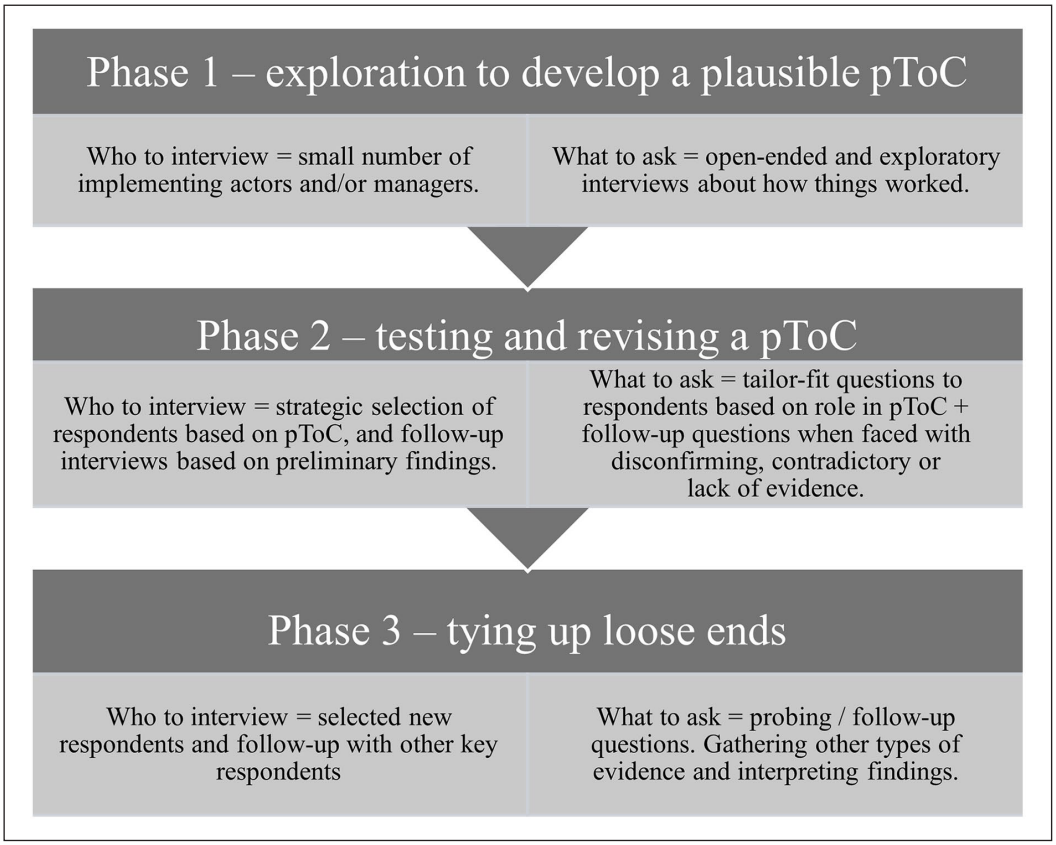

Figure 3. Interviews during different phases of PTEM.
Source. Authors.

in a process. The questions are crafted not to elicit theories from stakeholders about what could causally connect two actions, but rather to explore each action on itself and to collect what were the intentions and perception of the actors themselves when they interacted. The clear depiction of actions and interactions makes it easier for the evaluator to identify who would know about a particular action and to select them for our interviews. In the next section, we develop concrete guidelines for using a pToC as the central element of our interviews.

PToC-based interviewing—Practical guidelines

Like other TBEMs, the process of data collection in PTEM can be structured in three distinct phases (see Figure 3). Interviews in PTEM have different analytical functions in each phase, with who we are interviewing and what questions we ask differing. In the following, we walk through each phase, concentrating on the testing and revising phase because it is the core of a pToC-based evaluation.

To illustrate how these guidelines can be used, we use an example of how interviews were used in a PTEM evaluation of an AVRR program. The program aimed to support migrants to return and reintegrate into their country of origin. It did so by providing individual support to returning migrants to start a small business or find a job, and by developing capacities at the

institutional level to strengthen local public job agencies. Building on a re-constructed ToC of the program measures, the evaluation team developed pToCs for different causal pathways within this simple ToC, which included both individual level pToCs and institutional level ones. For the purpose of this illustration, we will focus on the institutional pToC, which concentrates on capacity development in the local public job agencies through a "job shadowing" intervention. The evaluation included case studies to assess effectiveness, impact, and sustainability of AVRR measures in three countries of origin. Data collection for the evaluation was conducted during each of the three phases.

Phase 1—Exploration to develop a plausible pToC

During the exploratory phase, the main purpose of interviews is to act as one source of empirical information related to what happened and the obstacles that the intervention had to overcome to achieve a contribution. The starting point of this can be a program theory or intervention logic if it exists in the program documents. How intervention logics can be developed into simple ToCs has been explained at length in TBE literature (Chen, 1990; Funnell and Rogers, 2011; Rogers and Weiss, 2007). Yet, for PTEM the exploration phase includes not only theorizing the overall linkage but also trying to unpack what the key challenges/barriers were and how they might plausibly have been overcome, that is, our key episodes. Ideas about what might have been done to overcome barriers can be developed using program documents together with logical brainstorming, relevant literature from the social sciences, other evaluations, but also interviews (for a practical guide on theorizing, see Camacho Garland and Beach, 2023).

Who to interview. In the exploratory phase, interviews would concentrate on people who can shed light on the challenges and obstacles and how the intervention might have overcome them. This can include independent experts, that is, well-informed academics, program managers, and project implementers. While we might interview several outside experts, it can make sense to be strategic regarding how many project implementers and program managers are interviewed in this phase given that we might only be able to interview them once. Therefore, waiting until we have more focused questions in the testing/revising phase can make sense. At the same time, to develop a pToC, it is typically necessary to interview a few key actors who can shed light on the process the intervention triggered. If possible, some of these key actors will be cultivated to act as confidants throughout the evaluation who can be returned to repeatedly with questions and inquiries about documents and the like.

What would we ask them about—Soaking and probing to build a pToC. Interviews in this phase are used to gather evidence to build the pToC, that is, determine the actual intervention and the possible contribution and how those two are linked in a causal pathway. In this phase, we want to gather information regarding:

1. Challenges the intervention sought to overcome to identify episodes.
2. Concrete actions performed during the intervention to overcome these challenges, and who performed them.
3. Plausible overall causal pathways that could link the intervention with the contribution.
4. Potential contextual conditions, that is, supporting and constraining factors that might be relevant for how the intervention worked.

Regarding the challenges, we want to ask about what usually happens in the context of this type of intervention. Questions can probe the magnitude of the contribution in terms of how big the challenges and barriers were that stood in the way of the contribution being produced. Second, the interview would ask implementers about the concrete actions within the program in terms of who did what in the key episodes. Program managers would be asked about their interpretations of the causal logic behind the sequencing of particular actions in the intervention. Third, explorative interviews would ask critically how the program should have/could have contributed to the outcomes, that is, discussing potential causal pathways. As with other TBEMs, questions about context will also be asked, that is, what are the factors that support or constraint the causal pathway for it to achieve the theorized contribution (Cartwright and Hardie, 2012: 61). Ideas about relevant contextual factors can also be drawn from existing social sciences literature, other evaluations, and our own logical reasoning about factors that might be relevant. Together, this information is used to inform questions about what might have been required for a particular action and response to happen in the theorized way in phase 2 (see below).

In our example of evaluating the AVRR program, we conducted initial interviews with experts in return and reintegration, several of the commissioners from the ministry, and program managers. First, a simple ToC of the AVRR program was re-constructed from program documents and the first round of probing interviews with stakeholders. The initial ToC was discussed with program managers and later taken to the exploratory case study. After validating the ToC, the team developed more disaggregated pToCs for the areas within the ToC which were expected to be most important for a causal explanation. In doing so, multiple pToCs both at the level of individual assistance to returnees and the level of institutional capacity development were established. In relation to the institutional pathway, program documents did not provide detailed information, but on-site interviews with program managers and staff of the advisory center and partner institutions helped shed light on the key episodes of the pToC. The respondents informed us that the part of the intervention that targeted this was the job shadowing program, in which officers from local public job agencies in the returnees' countries of origin worked in the donor-funded advisory center for a few months. This enabled the officers to learn about the needs of (return) migrants and acquire counseling techniques from more experienced advisers, at least in theory. Figure 4 depicts the preliminary pToC.

Phase 2—Testing and revising a pToC

In this phase, the main purpose of interviews is to explore whether the evidence expected by the preliminary pToC was present or not. When the interviews and additional evidence are not consistent with our pToC, we use the new information to revise the pToC. Since the pToC lies at the heart of the evaluation, deciding who to interview and crafting the questions would be based on it.

Who to interview. When evidencing a pToC, it is useful to differentiate between those who manage and implement a program, and those impacted by it (i.e. participants) (Adams, 2015: 494; Manzano, 2016: 349–350). Using the pToC developed in phase 1, we can further distinguish interview partners and talk to those performing actions, those impacted by them, and

more independent witnesses of what took place. Note that different actors might be relevant for different key episodes of the pToC. When deciding who to interview, we would want to consider who will make a better source (someone who worked in the program last year might have clearer recollections than someone who implemented it 5 years before) and who has less incentives to misrepresent (a program manager whose funding is on the line might be less forthcoming than one that is already moving on to another project).

The number of respondents to be selected depends on the type of program, and it might vary across different key episodes. In smaller scale and simpler programs, there might only be a handful of relevant respondents, meaning the selection would be focused on talking to the central actors in the intervention. As discussed in the section "Unpacking how an intervention works using a pToC," there might be many episodes of the pToC that play out in parallel; in effect, there might be many different "cases" of the episode. When there are many iterations of an episode that play out, some form of sampling strategy will have to be employed to ensure that our findings hold across the population affected by the intervention. How the episode plays out might be impacted by individual characteristics (e.g. the resources possessed by a participant might impact how they respond to the actions of program actors) and/or by contextual differences (e.g. part of a program implemented in a village with poor infrastructure might have different impacts than similar actions in another village with adequate infrastructure).

Note that representativity here is not statistical representativeness, but instead representativeness of the different causal dynamics that might be at play in an episode. Random sampling is not a good strategy when it is impossible to interview enough respondents to ensure a statistically representative sample. Instead, we would engage in a purposeful selection based on what individual and contextual factors might impact how the theorized episode (or episodes) play out. The number of respondents to be selected to evidence an episode therefore depends on the diversity of impacts and the contextual factors that might determine when one causal pathway is more likely than others. Note that the factors determining selection might be updated during the course of the evaluation as new information is gathered, but they will always be determined by the best way to gain evidence about our key episodes and our pToC.

If the episode involves program actor–participant interactions, and we are uncertain about what individual and contextual factors might be causally relevant, a good strategy is to select "success" and "failure" cases. Some iterations of the interactions might have produced good results, whereas others had little or even adverse effects. These cases could then be compared to detect what individual/contextual factors differed between the types of cases. It could be found that failure cases in a program involved participants with limited social networks, whereas the interaction produced positive results for participants with larger networks. Based on this, we would then deliberately select respondents from both groups to map out the different causal dynamics across different types of cases of the episode.

The selection of respondents should be open and flexible during the course of the evaluation as new empirical findings update our knowledge of how the intervention worked. We might find that key episodes worked in a different fashion, which would lead us to conduct follow-up interviews with new respondents, or we might want to return to interviewed respondents with new questions. Furthermore, if contradictory accounts are given, we would deploy additional follow-up interviews either with the same respondents or others

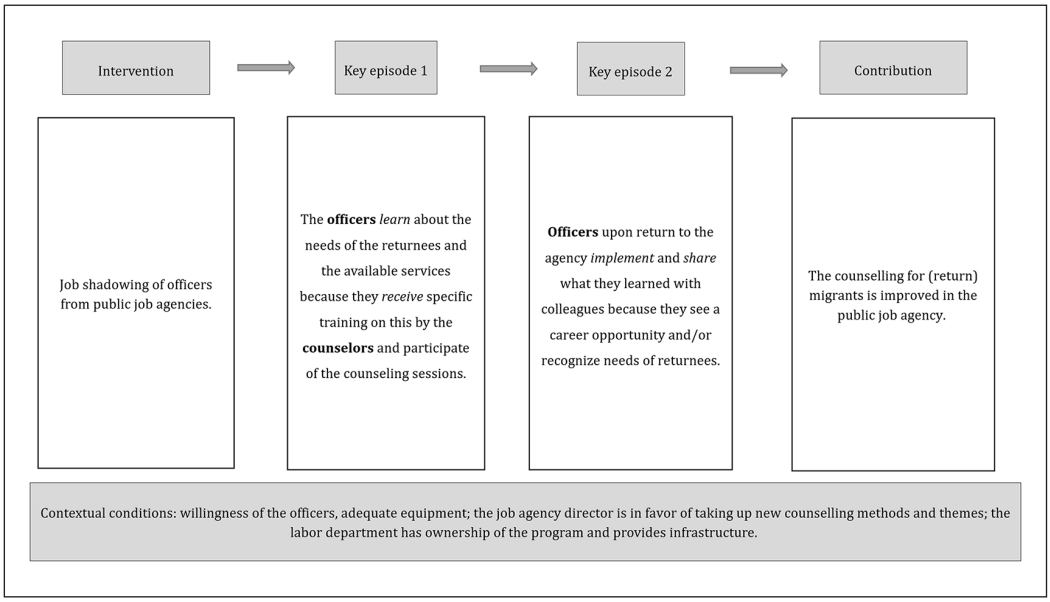

Figure 4. The institutional pToC: job shadowing.

in order to reconcile the findings. These interviews would loop back into our pToC, which would then be updated. If resources allow, the stopping rule for selecting new respondents would be some form of saturation (Saunders et al., 2018), that is, when further interviews do not tell us anything new and we are pretty certain that we have captured the diversity of the actor groups involved.

In our example, we applied the strategy of purposeful sampling and selected people for the key episodes of the pToC (see Figure 4). Both at the donor-funded advisory center and on the partner side, interviews were conducted at the management and the operational level. Covering the first key episode of the pToC, selecting directly involved actors was key. These were the counselors at the center and from the public job agencies that participated in the job shadowing. To explore the changes in the public job agencies after participation in the shadowing program (second key episode), interviews were also conducted with non-participating counselors within the selected agencies. To generate variance in important contextual factors (i.e. infrastructure and capacity of the job agency) across cases of the intervention, we selected one job agency within the capital city and one outside of it.

What to ask—Tailor-fitting topic guides to particular respondents. The topic guide should guide the interview. However, given that the purpose of the interviews is to critically assess the pToC, the interviewer should be flexible during the course of interviews. For instance, the respondent might provide a strong disconfirmation of a key episode which would lead us to inquire about how it then did work instead, with specific attention to the actions, interactions, and actors that compose it. The topic guide will be tailor-fit for different types of respondents based on what they can tell us about the workings of the pToC. Depending on the pToC under scrutiny, different actors might be active in very different parts of our pToC. Interviewees might be relevant

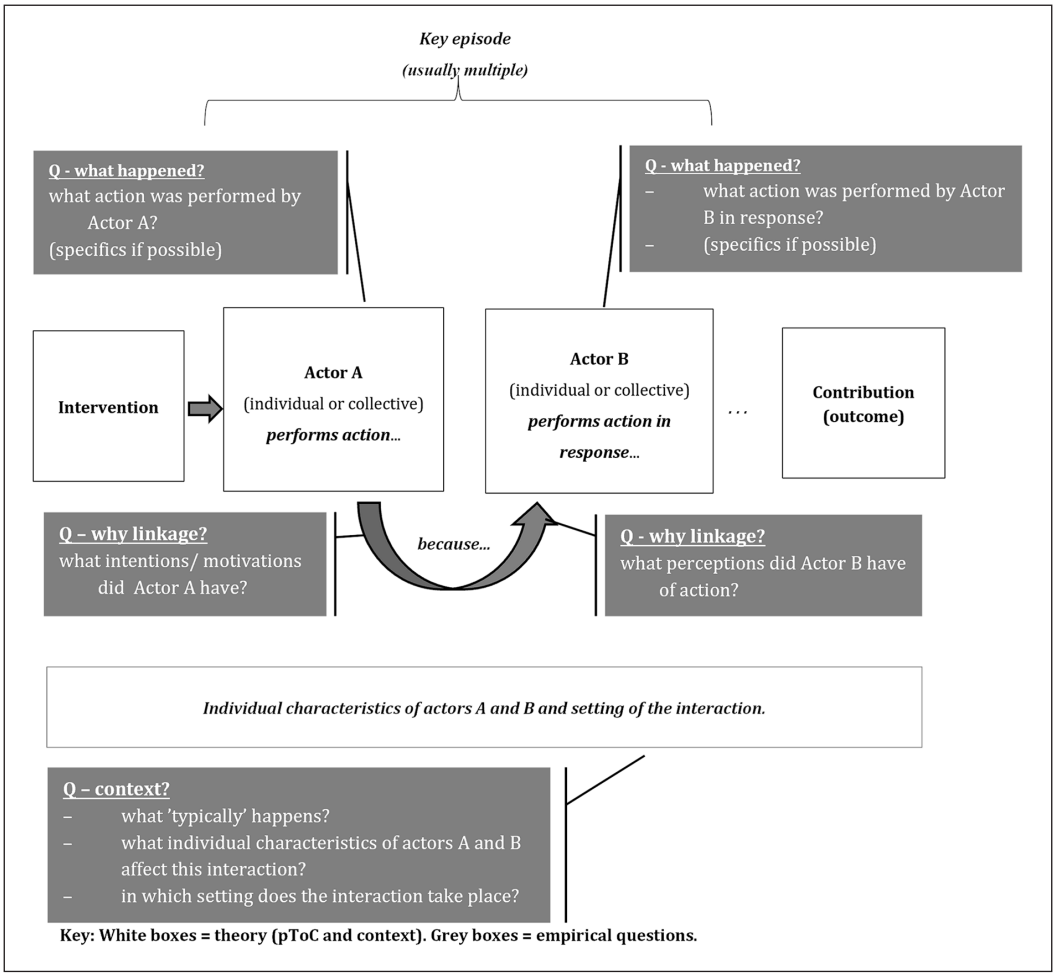

Figure 5. Interview questions when evidencing key episodes of a pToC.

informants for one key episode but not for another. The function of the particular interview is also important when crafting a topic guide. Is the respondent expected to give us an account of what took place, to corroborate or disconfirm other accounts, or to interpret other types of evidence (e.g. helping us to understand what a particular policy document means in the case context)? A topic guide for a central participant in a key episode might be quite lengthy, whereas a follow-up interview used only to corroborate another source might be very brief.

A final consideration is that respondents who were central in a process might also be able to provide other types of evidence (e.g. documents) that we otherwise cannot gain access to. If this is relevant, the topic guide should include a careful inquiry into whether it is possible to see particular documents (e.g. power point presentations made by an individual).

Figure 5 illustrates the types of questions we would be asking to test and revise a pToC. First, we must have an idea of *what* actions took place in the case. Some of this information might have been gathered during the exploratory phase, but most will be gathered during the

testing/revising phase. Mapping actions is focused on the initially identified key episodes, and whether the actions and responses they were expected to trigger took place as theorized.

To inquire about *why* actions took place in the case, we develop a set of questions that probe the motivations, intentions, and perceptions of involved actors in order to shed light on the linkages between actions and responses. To understand why actors do things, and why they respond in particular ways, we unpack the intentions, motivations, and meanings underlying actions from Actor A, and on the receiving end, how Actor B perceived the action and why it led to a particular action in response (Camacho Garland and Beach, 2023; Fujii, 2018; O'Mahoney, 2017). Questions that probe motivations and intentions would take the form of "do you recall why you did X," "what was your intention in doing X." Questions probing perceptions of actions can take the form "when A did X, how did you feel," "did you appreciate what A did," and "did you understand the information provided."

Linkages are not only at the individual level. Depending on the pToC, there might be interactions between meso-level actors (e.g. groups and organizations). For this type of linkage, a pToC would draw on broader social science theories (Camacho Garland and Beach, 2023). Social-psychological theories of in-out group dynamics, for instance, could be tested by interviewing persons from the "in" group asking questions that aim to explore how they thought about persons from the "out" group.

Questions about contextual factors would also be deployed. As mentioned previously, how a pToC and the episodes within it play out in particular cases depend on the individual characteristics of the actors—characteristics that typically go beyond socio-demographic differences, but instead are whatever factors might impact how an episode of interaction plays out, for example, cultural understandings, religious views, psychological well-being, and so on—and the setting in which their interaction took place—again broadly understood as anything that is causally relevant for how the interactions work, for example, the physical location, the political environment, the economic context, and so on. Put simply, what is considered relevant is determined by our pToC. Relevant factors are anything that can affect both sides of the action–response linkage, for example, they might affect how the action is received, or the possibility of another actor to respond in a particular way.

In PTEM, we need to theorize what context factors can make it more likely that the theorized pToC works as theorized (supporting factors) and which ones make it less likely (constraining factors). While some of the information we can ask directly, for example, to describe the setting in which the interaction took place or the budgetary constraints of the government, other information might be asked more indirectly, for example, asking about a participant's relationship with their family and community to understand their emotional well-being and societal status. It will be the role of the evaluator to assess this information to know what was necessary for the action to play the theorized causal role (Cartwright and Hardie, 2012). This would lead us to a refinement of the contextual relevant factors and to theorize about how they affect the causal pathway.

To assist in designing a topic guide, we can use a table where we note what we would like to know and how we would find out whether it happened (what empirics could we use) and what this means in terms of questions for the topic guide. Further building on the practical example on the evaluation of the AVRR program, Table 1 illustrates the interview questions for a selection of empirical fingerprints to probe a key episode of our pToC.

Table 1. An example of a topic guide for evidencing an episode of a pToC.

Key episode 1: "The officer learns about the needs of the returnees and the available services because they receive specific training on this by the counselors and participate of the counseling sessions."

Questions for interview with officer participating in the job shadowing (receiving end)	Questions for interview with the management level of the advisory center (hosting end)
• Did you receive any special training when you started? • Did you take part in any of the counseling sessions that the center has with returnees? • Were these counseling sessions useful? • Could you tell us more about what you think are the more pressing needs of returnees? • Was there any sort of structure or plan that guided the program?	• What are the components of the program and how are they being implemented? • Can you tell us about a typical counseling session with an officer shadowing a counselor? • Can you provide us with the plan that guided the program?

Source. Authors.

Assessment of evidence and follow-up interviews. Instead of assessing interviews only after fieldwork, PTEM involves a running assessment of what the conducted interviews tell us about a preliminary pToC. Given the ambition to update and revise the pToC during the course of an evaluation in PTEM, fieldwork would be adapted to what the interviews already conducted found. This requires that after interviews, the evaluator briefly assesses the evidence gained from interviews.

Assessing the core take-homes directly after each interview can be done using what we termed an "evidence grid." The evidence grid depicts the key episodes of our pToC in the columns, with the rows used for the key take-homes from each interview conducted. In Table 2, we illustrate the evidence grid for our institutional pToC. We also included an assessment of probative value for each interview where we highlighted a significant lack of knowledge or risk of not telling the truth.

In our evaluation example (see Table 1), we noted some risk of not telling the truth in interview 1 given the circumstance of the interview partner being in the position since the start of the program (highly knowledgeable) and in charge of setting up the program (some risk of not telling truth). An inverse situation shows for the informative value of interview 4, which had been conducted as a follow-up several months after interview 1. The interview partner just started the management position recently (less knowledgeable) and was not in charge of setting up the program being evaluated (less risk of not telling truth). Note that it is notoriously difficult to assess the underlying motivations of actors. Even if an actor does not have motives to embellish the truth, they might simply not remember correctly what happened; they might want to give you the answer they think you are expecting; or they might "overintellectualize" their answers. Given these challenges, we recommend that evaluators focus on corroboration across sources; in particular for the most crucial linkages in episodes. Finally, it is important to take into consideration that even if a respondent is not being completely forthcoming in their responses why they are not might nonetheless provide us with useful information (Fujii, 2010).

After we have conducted some interviews, the evaluator would begin thinking about the overall evidential picture. A step-by-step procedure is depicted in Figure 6. This preliminary

Table 2. An evidence grid for the institutional pToC.

	Intervention	Key episode 1	Key episode 2	Contribution	Informative value (knowledge + risk of not telling truth)
Interview 1 (hosting end, managing level)	Provides figures on number of officers participating over time	Reports of structure and work plan of the shadowing program			Very knowledgeable about the program; some risk of overstating
Interview 2 (receiving end, officer 1)	Reports participation in the shadowing program	Received specific trainings in the early phase of the shadowing program but program not structured	Reports of changing counseling techniques after return to the public job agency	Reports of positive feedback from migrant clients	Good account of memory; neutral description of program
Interview 3 (receiving end, officer 2)	Reports participation in the shadowing program	Reports in detail about the needs of returnees; describes the shadowing as not very structured	Describes peer learning on the job to transfer knowledge to colleagues	Self-reporting improved migrant counseling based on learned approach to counseling	Some inconsistencies; overly positive about the program and his role in peer learning (key episode 2)
Interview 4 (follow-up with hosting end, managing level)		Admitting that program curriculum is in progress			Less knowledgeable about the program; less risk of not telling truth

Source. Authors.

assessment involves weighing the collected evidence—of which interviews are only a part—for each key episode. When there are multiple cases of a key episode, we are also assessing whether there are important differences in how the episode played out across the cases, as well as potentially contradictory evidence within any given case.

A key episode is confirmed to some degree when we have some form of evidence that the actions and responses took place as theorized, and we have some evidence of the linkages. This typically requires multiple pieces of evidence from different sources. There are often two sides to a story, meaning that if we are evidencing interactions between program actors and beneficiaries, we would want to find similar accounts from both sides, and/or corroboration from other types of evidence. This means both assessing whether the perspectives from the different types of actors are compatible in terms of our pToC having taken place as depicted (and dig into the places where it is not), and whether they are internally consistent (i.e. the same type of actor is giving us relatively similar information regarding what happened). For example, it would be problematic if the management of the center tells us every participating

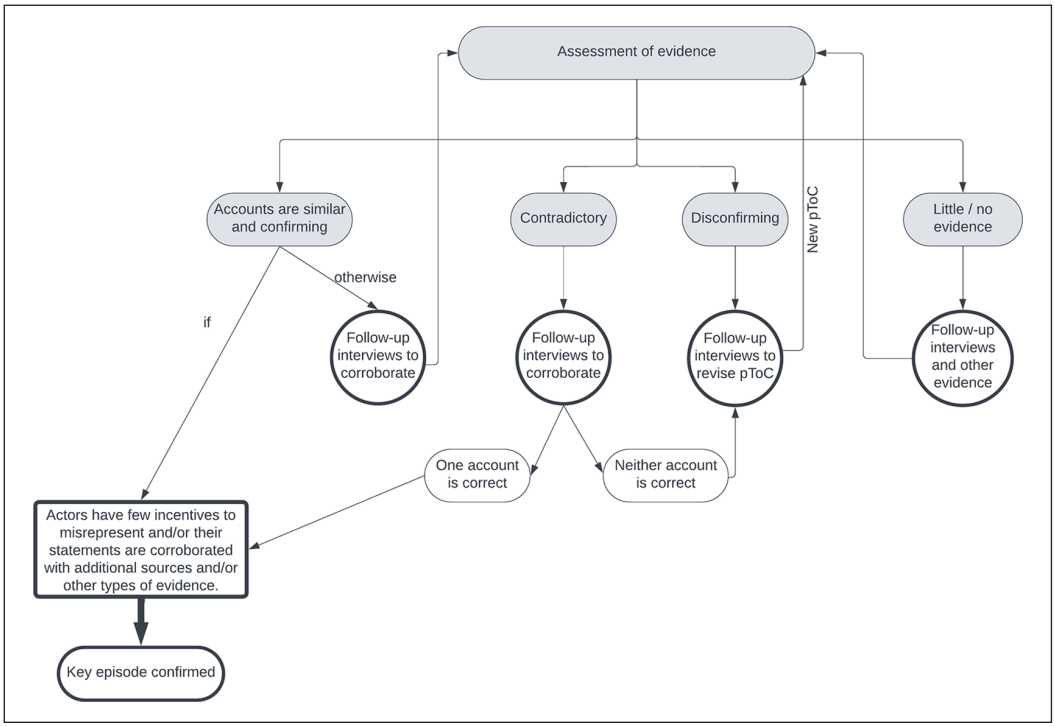

Figure 6. Assessment of evidence.
Source. Authors.

officer needed to go through a 1-week capacitation on how to counsel but neither of the officers remembers going through one. This would require further probing. We could then learn that the training had only begun being implemented, and the officers had taken part of the job shadowing before this, or it could be that although the training was designed, there were not enough staff resources to allocate to it and thus it was never implemented.

It is important to keep in mind that confirmation or disconfirmation is more a continuum than a dichotomy and we will never find evidence that confirms or disconfirms with 100 percent certainty a particular pToC (Befani, 2021). Instead, we try to assess in what direction the aggregate evidence points. If our sources are weak, we could try to collect more information and other evidence types. Ideally, even strong interviews would be corroborated with other types of evidence and/or other interviews.[1] However, in many evaluation situations, there will be episodes that cannot be evidenced strongly. When this happens, the evaluation should clearly flag that there is only weak evidence behind this part of the contribution claim.

If contradictory evidence is found, we would (if possible) conduct follow-up interviews to figure out what might have actually happened. It is important to consider that contradictory accounts do not necessarily mean one party is lying. Instead it could also point to reality being more complex than expected or someone not recollecting events perfectly. To reconcile the evidential picture, we would want to ask what could explain this clash of information, drawing from the theory and considering other observable measures. Using our AVRR evaluation example, we found contradictory evidence on the nature of the shadowing program. Statements on the program's structure differed between the management level (sending end) and the

Table 3. A code tree for the institutional pToC.

Generic code	Sub-code
Intervention	Infrastructure support
	Shadowing program implementation
Key episode 1	Officer learning about needs of returnees
	Officer learning about services for returnees
	Other learnings
Key episode 2	Officer uses lessons learned in counseling work
	Officer shares lessons learned
	Officer implements other lessons learned
Contribution	Counseling for (return) migrants is improved
	Changes in the working system are implemented
	Migration desks are created in job agencies
Context factors	Supporting factors
	Constraining factors

Source. Authors.

participating officers (receiving end). This contradiction could be resolved by conducting a follow-up interview with the new management at the center some months after the first interview. Furthermore, by asking for additional information (i.e. the shadowing program's work plan) from the donor organization, we could fully reconcile contradicting statements from the initial interviews.

In the case of finding disconfirming evidence, we would want to revise our pToC. Disconfirming evidence can be regarding actions which have not taken place as planned by the program or regarding the linkage between intentions and perceptions. In both cases, we would want to figure out what happened and why, as with the previous example. At any case, when disconfirming evidence arises, we would revise our key episode and develop a new one that we can then test, or explore what went wrong. It could be that the implementation did not work as expected/planned and/or that the contextual factors that are needed for the causal linkage to work as expected were not present.

If little or no evidence is found for a key episode, the lack of evidence can mean two different things: either the key episode did not take place (disconfirmation), or it might have but we simply do not have access to evidence that would allow us to assess this. Absence of evidence is evidence that the key episode did not take place when the actions/responses should have left empirical fingerprints, but we do not find them. In the second instance, we would engage in follow-up interviews with new respondents and/or search for other types of evidence. For instance, if we do not have access to participants, we might consider other more indirect and circumstantial evidence. For example, could there be patterns in the documentary record that suggest that an action took place at a particular moment in the process? An actor might have presented an idea in a meeting, after which the idea is reflected in subsequent documents, whereas it was not present before.

Phase 3—Documenting and tying up loose ends

Finally, after fieldwork, the evaluation turns to a more systematic assessment of what was found. Do we have enough confirming evidence or does more evidence need to be collected? Are

further revisions to the pToC required? The assessment will hardly be a pass or fail test, but rather a matter of degrees of confidence based on the strength of confirming evidence gathered. If we still have not received relevant documents, this is the time to follow up on those.

If the interviews are coded, codes would be directly related to different aspects of the pToC. In our example, coding of the different interviews was made using MAXQDA, and while some codes were created to address other evaluation questions beyond the specific contribution (e.g. the nature of the cooperation between different actors), one part of it closely followed the key episodes of our pToC. The code tree for the pToC is depicted in Table 3 (we have adapted the codes for comprehension purposes, the original ones are shorter and use acronyms to facilitate their use in the software).

During this phase, the empirical material will go through a more systematic analysis, both individually and in a collective sense, repeating the assessment with data for each key episode and the whole pToC. This is important because at this stage we have more time to look at the overall evidential picture across episodes. If we were dealing with multiple cases for one or more episodes, now is when we assess what we learn from each of them and aggregate them together. We cannot confirm a pToC if it was present in only one case. If a number of cases show evidence that the linkage worked as expected, we will not be able to claim that it works on average, but we can provide evidence of how it worked in our cases and what context factors enable or constrain the causal pathway. If the expected contribution did not occur in most cases, we would want to compare the cases where it worked and did not detect the contextual factors that needed to be present. The evidential threshold required depends on the purposes of the evaluation. If an evaluation is to be used for accountability purposes, very strong confirming evidence for how the pToC actually worked should be attained. In contrast, an evaluation of a project pilot might only require that some confirming evidence that makes it more plausible that the intervention worked as expected.

The evaluation might find that elements of the pToC were not as important as initially hypothesized. The evidence might show that a key episode was not key in light of another key episode or in relationship to the causal pathway. If this occurs, the evaluator should revise the theory, and probably also return to some sources for further information. It could also be that although we did not have enough strong evidence for a particular key episode we have strong evidence for the others, and the evidential threshold of our evaluation might still be met.

We would want to come back to the depiction of our pToC. The key episodes of our pToC are parts where evidence is key to get a fuller picture and thus see where we have enough, where do we need more, where are we uncertain, where would we need to go back to our data, and so on. If after this more systematic assessment you consider that the pToC has weak evidence (and it will usually be that some key episodes of our pToC have stronger evidence than others), there are two possible scenarios. One, you reach more cautious conclusions, highlighting that a key episode of the process is in a black box. Two, if the evidence for the rest of the pToC makes more plausible how a key episode with weaker evidence works, it is ok to speculate, but this would be flagged.

Conclusion

As with other TBEMs, PTEM puts the theory at the center of the evaluation. In the case of interviews, this means that the respondents are selected and the questions crafted based on the pToC and will be used differently in three different phases of the evaluation. Using the pToC

to structure and assess the evidence during the process of data collection helps (a) to close evidence gaps and reach a saturation of evidence and (b) to communicate results to the evaluation's stakeholders in a timely and efficient manner.

In the exploration phase, interviews will be used to build a pToC of the intervention and possible contribution. Here, the use of PT interviewing helps to specify pToC, where the causal linkage is clearly spelled out, enhancing the causal element of TBEM. In the testing and revising a pToC phase, interviews will serve to assess whether the key episodes took place as depicted and to revise out pToC when they did not. This technique allows for a more efficient and targeted data collection, where we can immediately know what we still need to find and thus inform what we ask in posterior interviews. Analyzing the evidence in the field does not only enable us to make timely changes to collect more information but also allows us to communicate preliminary findings quickly to stakeholders. In the final phase, we come back to the depiction of our pToC and see if we have collected enough evidence to make claims regarding its workings (to either claim it works as described or why it does not work as expected). At this phase, the data analysis focusing on the pToC helps us not to drown in data, identify evidence gaps, and tie up loose ends efficiently. Using disaggregated pToCs helps us keep in mind what are the evidence gaps and what information we still need to collect. This is particularly important in critical results settings, where assessing evidence along the pToC allows us to be more certain about the results and use evidence accordingly. Similar to other TBEMs, interviewing in PTEM requires time and effort. Structuring data collection along our pToC enables us to conduct more targeted case studies, improve the rigor when making causal inferences based on evidence, and communicate results quickly.

Acknowledgements

We would like to extend our gratitude to all those who contributed to the evaluation in which we piloted and further developed the use of process tracing in interviews. We sincerely thank Lena Heller, Lea Jechel, Felix Leßke, and Martin Bruder from DEval for all of their insights, and our consultants in the field, Harriet Osei-Boakye and Gloria Asante. Their contributions were instrumental. Any errors or oversights in this paper are solely the responsibility of the authors and do not reflect on the contributions or efforts of others involved in this work.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Independent Research Fund, Denmark (grant number 0133-00115B). The fieldwork that informed this paper was done in the context of an evaluation conducted by the German Institute for Development Evaluation (DEval), funded by the German Ministry of Economic Cooperation and Development (BMZ).

ORCID iDs

Gabriela Camacho Garland https://orcid.org/0009-0008-1826-932X
Derek Beach https://orcid.org/0000-0003-1869-0145

Note

1. A note on triangulation. We use the term corroboration in relation to mixed pictures or accounts because we do not have a known independent source that we can use to assess what happened (action leaves traces, intentions and motivations put as in the world of people's lived truths). The term triangulation is somewhat misleading because we do not have a known, independent point of reference that we can use to gauge bias in other sources.

References

Adams WC (2015) Conducting semi-structured interviews. In: Newcomer KE, Hatry HP and Wholey JS (eds) *Handbook of Practical Program Evaluation*. Hoboken, NJ: John Wiley & Sons, 492–505.

Befani B (2021) *Credible explanations of development outcomes: Improving quality and rigour with Bayesian theory-based evaluation*. EBA Report. Stockholm: The Expert Group for Aid Studies (EBA).

Camacho Garland G and Beach D (2023) Theorizing how interventions work in evaluation: Process-tracing methods and theorizing process theories of change. *Evaluation* 29(4): 390–409.

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

Chen H (1990) *Theory-Driven Evaluations*. Thousand Oaks, CA: Sage.

Fujii LA (2010) Shades of truth and lies: Interpreting testimonies of war and violence. *Journal of Peace Research* 47(2): 231–41.

Fujii LA (2018) *Interviewing in Social Science Research: A Relational Approach*. New York: Routledge.

Funnell SC and Rogers PJ (2011) *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. San Francisco, CA: Jossey Bass.

Guion L, Diehl D and McDonald D (2011) *Conducting an In-Depth Interview* (FCS6012). Gainesville, FL: EDIS.

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

Mayne J (2012) Contribution analysis: Coming of age? *Evaluation* 18(3): 270–80.

Mayne J (2015) Useful theory of change models. *Canadian Journal of Program Evaluation* 30(2): 119–42.

Mayne J (2019) A brief on contribution analysis: Principles and concepts. Available at: https://repository.uneca.org/bitstream/handle/10855/49332/Chapter%2018_%28resource%20doc%29%2014pp%2C%20BriefonCA.pdf?sequence=382&isAllowed=y (accessed 29 September 2024).

O'Mahoney J (2017) Making the real: Rhetorical adduction and the Bangladeshi Liberation War. *International Organization* 71(2): 317–48.

Pawson R (1996) Theorizing the interview. *The British Journal of Sociology* 47(2): 295–314.

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

Pawson R and Tilley N (2004) *Realist Evaluation. Online Paper Funded by the British Cabinet Office*. London: British Cabinet Office. Available at: https://www.researchgate.net/publication/264082581_Realist_Evaluation (accessed 28 July 2023).

Rogers PJ and Weiss CH (2007) Theory-based evaluation: Reflections ten years on: Theory-based evaluation: Past, present, and future. *New Directions for Evaluation* 114: 63–81.

Saunders B, Sim J, Kingstone T, et al. (2018) Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity* 52: 1893–907.

Schmitt J and Beach D (2015) The contribution of process tracing to theory-based evaluations of complex aid instruments. *Evaluation* 21: 429–47.

USAID (2011) *Conducting Key Informant Interviews* (TIPS 2). Washington, DC: USAID.

Wauters B and Beach D (2018) Process tracing and congruence analysis to support theory-based impact evaluation. *Evaluation* 24: 284–305.

Gabriela Camacho Garland was a postdoc at the University of Aarhus, Denmark and guest researcher at the German Institute for Development Evaluation (DEval) during this research. She works on the application of process tracing methods in policy and development evaluation. She is currently a Research Coordinator at Transparency International.

Derek Beach is a Professor of Political Science at the University of Aarhus, Denmark. He has authored articles, chapters, and books on process tracing methods, policy evaluation, and the role of expertise in policy processes in the European Union. He has co-authored the book Process-tracing Methods: Foundations and Guidelines.

Johannes Schmitt is a Senior Evaluator at the German Institute for Development Evaluation (DEval), where he currently leads an evaluation on Return and Reintegration Assistance. He has worked conceptually and practically with theory-based evaluation methods, including process tracing.