9 Process Tracing for Program Evaluation

Andrew Bennett

9.1 Introduction

In recent years, the "replication crisis," or the finding that many attempts to replicate prominent published studies have failed to reproduce their original results, has roiled the medical, social science, public policy, and development research communities (Ioannidis 2005).[1] This has led to efforts to change both procedures and cultures in carrying out and publishing research, including a de-emphasis of p-values in statistical research, preregistration of studies using experimental designs or observational statistics, and, in some journals, preacceptance of studies based on their designs rather than their results.

Although many of the projects whose results could not be replicated were experimental studies, one response in the program evaluation community has been to increase the emphasis on experiments. Done well, these research designs, including field experiments and natural experiments as well as lab and survey experiments, remain powerful tools in program evaluation. Yet experiments impose demanding methodological requirements (Cook 2018; Deaton and Cartwright 2018), they face challenges of external validity, and in some policy domains they are not practical for fiscal or ethical reasons. In addition, evaluators are often called upon to evaluate programs that were not set up as experiments, including programs instituted quickly to address pressing needs.

[1] It is important to note that failure to replicate a study's findings does not necessarily mean the study's results are false; some studies cannot be replicated, for example, because it is no longer possible to replicate their particular context or sample.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

Thus, a second response in the evaluation community has been increased interest in “process tracing,” a method of causal inference that is applicable to single observational case studies (Bamanyaki and Holvoet 2016; Barnett and Munslow 2014; Befani and Mayne 2014; Befani and Stedman-Bryce 2017; Busetti and Dente 2017; Mendoza and Woolcock 2014; Punton and Welle 2015; Schmitt and Beach 2015; Stern et al. 2012; Wauters and Beach 2018). Process tracing has been common in political science for decades and has been the subject of recent methodological innovations, most notably the explicit use of Bayesian logic in making inferences about the alternative explanations for the outcomes of cases. Process tracing and program evaluation, or contribution analysis, have much in common, as they both involve causal inference on alternative explanations for the outcome of a single case (although process tracing can be combined with case comparisons as well). Evaluators are often interested in whether one particular explanation – the implicit or explicit theory of change behind a program – accounts for the outcome. Yet they still need to consider whether exogenous nonprogram factors (such as macroeconomic developments) account for the outcome, whether the program generated the outcome through some process other than the theory of change, and whether the program had additional or unintended consequences, either good or bad. Process tracing can address these questions, and it is also useful in assessing the validity of the assumptions behind natural, field, and lab experiments.

This chapter outlines the logic of process tracing and the ways in which it can be useful in program evaluation. It begins with a short discussion of the philosophy of science underlying process tracing and a definition of process tracing. It then turns to the role of process tracing in single case studies and in checking the underlying assumptions of experiments, field experiments, and natural experiments. Next, the chapter provides practical advice on process tracing for causal inference in individual cases and discusses the special considerations that arise in the use of process tracing in program evaluation. Finally, the chapter outlines an important recent development in process tracing methods: the explicit and transparent application of Bayesian logic to process tracing. It concludes that explicit Bayesian process tracing holds promise, but not yet proof, of improving the use of process tracing in causal inference and program evaluation.

9.2 The Philosophy of Science of Causal Mechanisms and Process Tracing

The increased interest in process tracing across the social and policy sciences is related to the turn in the philosophy of science over the last few decades

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

toward a focus on causal mechanisms as the locus of causal explanation. Earlier, philosophers hoped that either "laws" or observed relations of statistical conditional dependence – analogous to what the philosopher David Hume called "constant conjunction" – would provide satisfactory accounts of causation and causal inference. The attempt to explain outcomes by reference to "laws" or "covering laws" foundered, however, when its advocates, including Carl Hempel, failed to come up with a justification or warrant for laws themselves (Salmon 1998, 69). In addition, Hempel's approach, known as the "Deductive-Nomological (D-N) Model," had difficulty distinguishing between causal and accidental regularities. In a common example, a barometer's readings move up and down with changes in the weather, but they do not cause the weather. Rather, changes in air pressure, which are measured by a barometer, combine with changes in temperature and other factors (topography, humidity, ocean currents, etc.) to cause the weather. But the D-N model has trouble distinguishing between a barometer and a causal explanation of the weather, as the barometer readings exhibit strong law-like correlations with the weather.

In an effort to address these problems, philosopher of science Wesley Salmon attempted to work out a defensible schema of explanation based on conditional dependence, or, in Salmon's terms, "statistical relevance." After encountering several paradoxes and dead-ends in this effort, he ultimately concluded that "statistical relevance relations, in and of themselves, have no explanatory force. They have significance for scientific explanation only insofar as they provide evidence for causal relations . . . causal explanation, I argued, must appeal to such mechanisms as causal propagation and causal interactions, which are not explicated in statistical terms" (Salmon 2006, 166).

Many philosophers and social and other scientists thus turned to exploring the role of causal mechanisms and causal processes in causal explanation and the roles of different research methods (experiments, observational statistics, case studies, etc.) in uncovering evidence about the ways in which causal mechanisms work and the contexts in which they do and do not operate. Within philosophy, the discussion of causal mechanisms has generally gone under the label of "scientific realism" (related but not necessarily identical approaches include "causal realism" and "critical realism"). This is the school of thought that Ray Pawson, Nick Tilley, and others in the evaluation community have drawn upon in their discussions of "realist evaluation" (Astbury and Leeuw 2010; Dalkin et al. 2015; Pawson and Tilley 1997).

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

A detailed analysis of scientific realism and causal mechanisms, and of debates surrounding their definitions, is beyond the scope of the present chapter, but a brief summary will suffice. Realism argues that there is an ontological world independent of the mind of the observer or scientist, and causal mechanisms ultimately reside in that ontological world. Scientists have theories about how causal mechanisms work, and, to the extent that those theories are accurate, they can explain outcomes. In one widely cited formulation, causal mechanisms are "entities and activities, organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer, Darden, and Craver 2000, 3). In another definition that also emphasizes a kind of regularity, mechanisms are processes that cannot be "turned off" through an intervention (Waldner 2012). Fire happens, for example, whenever there is combustible material, oxygen, and a sufficient ignition temperature; we can intervene on the presence of oxygen or materials or the temperature, but we cannot intervene on the mechanism of fire happening when the requisite materials and conditions exist.

Mechanisms are in the world, and theories about mechanisms are cognitive or social constructs in our heads. Scientists make inferences about the accuracy and explanatory power of theories about mechanisms by outlining the observable implications of these theories and testing them against evidence. In frequentist studies, the observable implications of theorized mechanisms lie at the population level, such as the correlations one would expect to find if a theory were true. In contrast, process tracing gets closer to mechanisms where they actually operate: in individual cases. The operation and interaction of causal mechanisms is realized in specific cases and contexts, and scientists and evaluators are interested in building theoretical understandings of the conditions under which mechanisms are activated or deactivated and the ways in which they interact with other mechanisms.

In studying individual cases, process tracers focus not just on the values of the independent and dependent variables, but on diagnostic evidence of sequences and processes that lie in the temporal space between the independent variables and the observed outcome. Process tracing uses this evidence to make inferences about which theories most likely offer true explanations of a case's outcome, sometimes called "inference to the best explanation." Process tracers continually ask "What should be true about the sequence of events between the independent variables and the dependent variable if a theory is a true explanation of the outcome of a case?" In the social sciences, this often takes the form of asking "Who should have

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

conveyed what information to whom, when, and with what effect at each stage in the process if this explanation is true?" Diagnostic evidence, ideally, is information that allows inferences about which processes are in operation, but that does not itself represent an additional variable that independently affects the operation or outcome of these processes. Diagnostic evidence, in other words, is not an "intervening variable" in a process, as the term "variable" implies an independent entity with its own potential causal effects.

9.3 Definition of Process Tracing

Process tracing is "the analysis of evidence on processes, sequences, and conjunctures of events within a case for the purposes of either developing or testing hypotheses about causal mechanisms that might causally explain the case" (Bennett and Checkel 2015, 7).[2] Process tracing is a within-case form of analysis: that is, it seeks to explain the outcomes of individual cases (sometimes called "historical explanation" or "token explanation"). At the same time, process tracing can be combined with cross-case comparisons or other methods. Researchers can use process tracing, for example, to assess whether differences between most-similar cases might account for these cases' different outcomes. The theoretical explanations of case outcomes assessed through process tracing can be about individual mechanisms or processes, or combinations of mechanisms and processes. They can include structural mechanisms, agent-based mechanisms, or any combinations thereof.

A key difference between process tracing and frequentist statistical analysis is that statistical analysis faces the "ecological inference" problem: even if a statistical correlation correctly captures an average causal effect for a population, it does not necessarily explain the outcome for any particular case in that population. Process tracing, in contrast, focuses directly on the causal explanation of individual cases. It may or may not uncover strong evidence leading to a confident explanation of a case, but it does aspire to develop directly the strongest explanation of the case that the evidence allows. Rather than facing an ecological inference problem, process tracing explanations, even when strong, face challenges regarding the external validity or generalizability of findings from individual cases. As Chapter 4 argues, the challenges of generalizing the results of case studies, while real, are often

[2] The term "causal chain" analysis refers to methods quite similar to process tracing.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

misunderstood. The explanation of an individual case can indeed prove generalizable: a new understanding of how a causal mechanism works, derived from the study of an individual case, can give strong clues about the scope conditions in which we should expect that mechanism to operate.

While process tracing is most often focused on the explanation of case study outcomes, the logic of process tracing can also be used in interrogating the validity of the strong assumptions necessary for experiments, field experiments, and natural experiments. In lab experiments, in addition to carrying out various balance tests on the treatment and control groups, researchers can use process tracing to check the procedures through which individuals were assigned to one group or the other, to assess the ways in which and reasons for which individuals opted to drop out of one group or the other, and to check on the possible presence of unmeasured confounders. Similarly, in field and natural experiments, where there is less control over assignment to treatment and control groups, researchers can use process tracing to assess whether the actual assignment or election into treatment and control groups was "as if random," and to evaluate evidence on whether the hypothesized process does indeed account for differences between the outcomes of the treatment and control groups (Dunning 2015).

Process tracing is much like detective work: the researcher is seeking an explanation of one case, and they can use both deductive and inductive inferences to find the best explanation. Deductively, the researcher starts with some "suspects" – the theories that have typically been applied to the outcome of interest. In program evaluation, this includes the theory of change explicitly or implicitly adopted by a program's designers and managers, but it also includes alternative explanations that relate to variables exogenous to the theory of change, such as macroeconomic trends, demographic change, local and national political developments, wars, natural disasters, etc. The researcher then looks for evidence on the deductively derived observable implications of each potential theoretical explanation of the outcome of the case. Just as a detective can reason forward from suspects and backward from a crime to connect possible causes and consequences, researchers can trace processes in both directions. A researcher can trace sequences forward from the independent variables, asking whether each caused the next step in the hypothesized chain leading to the outcome, and the step after that, and so on to the outcome. She or he can also trace backward from the outcome, asking about the most proximate step in the process that caused the outcome, and the step prior to that, back to the independent variables.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

Deductively derived implications of a theory are one type of "clue," but researchers also gather other kinds of evidence or clues that they stumble upon inductively as they investigate or "soak and poke" in their cases. Inductively discovered evidence might point to an existing social science theory that the researcher had not identified as a possible explanation of the case, or it may lead to the development of an entirely new theory as a potential explanation of the case. It is possible that an inductively identified piece of evidence, even evidence for an entirely new theory or explanation, can be so strong – so uniquely consistent with one explanation and so inconsistent with all other explanations – that this theory could become the most likely explanation for the outcome even without further corroboration. This cuts against the common but erroneous intuition that a theory developed from a case can never be considered to have undergone a severe test from the evidence that led to the theory. Anyone who has done their own amateur home or car repairs knows the experience of finding physical evidence that not only suggests but makes highly likely a heretofore untheorized explanation for why a switch, appliance, or part is not working.

In addition, our confidence in a newly derived or newly added potential explanation of a case can be strengthened if the explanation entails additional observable implications within the same case that are then corroborated by additional evidence. This contravenes the frequent claim that one cannot develop a theory from a case and test it against the same case. We can develop a theory from a case and test it against *different evidence* from the same case that is independent of the evidence that gave rise to the theory. It would be illogical, for example, for a doctor to diagnose a rare illness in a patient based on an unexpected test result, and then insist on testing the diagnosis on a different patient, rather than on an additional diagnostic test in the first patient.

9.4 Practical Advice on Traditional Process Tracing

The general approach of process tracing is fairly intuitive as it follows a kind of inferential process that has been around as long as humankind. Yet despite its seeming simplicity and familiarity, researchers do not always do process tracing well, and, as the final section of this chapter argues, even trained researchers make common mistakes in employing the Bayesian logic that underlies process tracing. So how can we do process tracing well? Elsewhere I have elaborated with my co-author Jeffrey Checkel on ten best practices for

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

Box 9.1 Best practices in process tracing

1. Cast the net widely for alternative explanations.
2. Be equally tough on the alternative explanations.
3. Consider the potential biases of evidentiary sources.
4. Take into account whether the case is most or least likely for alternative explanations.
5. Make a justifiable decision on when to start.
6. Be relentless in gathering diverse and relevant evidence, but make a justifiable decision on when to stop.
7. Combine process tracing with case comparisons when useful for the research goal and feasible.
8. Be open to inductive insights.
9. Use deduction to ask "If the explanation is true, what will be the specific process leading to the outcome?"
10. Remember that conclusive process tracing is good, but not all good process tracing is conclusive.

Source: Bennett and Checkel (2015)

being a good traditional process-tracing detective; here, I introduce these practices briefly and elaborate on the considerations of each that are most relevant to program evaluation (Bennett and Checkel 2013, 20–31). In the final section of this chapter I address how to carry out the more formal Bayesian variant of process tracing.

9.4.1 Cast the Net Widely for Alternative Explanations

One of the most common mistakes in case study research designs is the omission of a potentially viable explanation. It is important to consider a wide range of potential explanations, as the omission of a viable explanation can skew the interpretation of evidence on all the other explanations that a researcher does consider. Explanations for program outcomes need not be – and usually should not be – single-variable explanations. Rather, they can include combinations of interacting variables. There are four main sources of potential alternative explanations of program outcomes. The first is the program's explicit or implicit theory of change, which should be evident in program documents and interviews with program managers. In practice, individuals may differ in how they view the theory of change or interpret its

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

implications for how they administer the program, so it may be necessary to process trace different variants of the theory of change. As it is essential to not unduly privilege the theory of change, a second source of explanations includes those offered by other stakeholders (beneficiaries, government officials, members of communities who experience knock-on effects, etc.), as well as the implicit or explicit explanations news reporters give for program outcomes. A third range of candidate explanations consists of social science theories that researchers have typically applied to the kind of program or outcome in question. As there is a wide range of such theories, a useful checklist is to consider both explanations focused on variations among agents (their interests, capacities and resources, networks, ideas, etc.) and those focused on social structures (norms, institutional rules and transactions costs, and actors' relative material resources).[3] Fourth, it is useful to consider the standard list of potential confounding explanations for program outcomes and to do process tracing on any that are relevant. These include:[4]

History: exogenous events (economic cycles, elections, natural disasters, wars, etc.) during the program period that can affect outcomes.

Maturation: program beneficiaries might go through aging processes that improve or degrade outcomes over time.

Instrumentation: changes in measurement instruments or technologies during the program can affect the assessment of outcomes.

Testing: exposure to testing or assessment can change the behavior of stakeholders.

Mortality: there may be selection bias regarding which stakeholders or recipients drop out of the program.

Sequencing: the order in which program treatments are implemented may affect outcomes.

Selection: if acceptance into the program is not random – for example, if the program chooses to address the easiest cases first (low-hanging fruit) or the hardest cases first (triage), there can be selection bias.

Diffusion: if stakeholders interact with each other due to the program, this can affect results.

Design contamination: competition among stakeholders can affect outcomes; those not selected as beneficiaries might try harder to improve

[3] For a taxonomy of twelve common types of social science theories based on different types of agentic and structural interactions, and approaches to explanations focused on material power, institutional transactions costs, and ideas and social relations, see Bennett (2013).

[4] Many of these are discussed in Shadish, Cook, and Campbell (2002).

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

their own outcomes, or they might become demoralized and not try as hard to succeed.

Multiple treatments: if governments or other organizations are administering programs targeted at similar outcomes, or if the program being evaluated includes multiple treatments, this can affect outcomes.

There can also be potential interactions among these factors that merit process tracing.

9.4.2 Be Equally Tough on the Alternative Explanations

Being fair to alternative explanations is an obvious goal for evaluation and causal inference, but it can be difficult to achieve in practice given the cognitive propensity for confirmation bias. A key contribution of rigorous research methods, whether qualitative, quantitative, or experimental, is to make it harder to engage in the well-known heuristics and biases through which individuals often make faulty inferences. Process tracing methods aim to achieve this by requiring that we consider not only what evidence would be consistent with each explanation, but also what other explanations might be equally or more consistent with that same evidence. They also require that we consider what evidence would be inconsistent with each explanation, and the degree to which other explanations would be (in)consistent with that evidence. This can prevent the temptation to focus mostly on affirming evidence for one explanation and to neglect how that same evidence could also fit other explanations. A common mistake occurs when researchers do deep process tracing on one theory, such as the theory of change, and only cursory process tracing on alternative explanations. An unbiased estimate of how likely it is that a theory is a good explanation of the outcome of a case requires that the alternative explanations receive scrutiny as well. Process tracing proceeds not only by finding evidence that fits one explanation better than the others, but also by eliminative induction of alternative explanations that do not fit the evidence. The discussion of Bayesianism in Section 9.5 gives a more formal assessment of how the relative likelihood of evidence given alternative explanations should affect the confidence we invest in those explanations.

9.4.3 Consider the Potential Biases of Evidentiary Sources

The potential biases of stakeholders are sometimes fairly clear, but they can depend on institutional and contextual factors. A government official might

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

want to cast a program in a good or bad light, for example, depending on their party affiliation. Program managers generally want to show that their program is succeeding, but they might be tempted to downplay the baseline achievements they inherited from their predecessors. It is important as well to consider not only motivated biases, but also unmotivated biases that can arise from the selective information streams to which individuals are exposed, or from procedures through which some documents are maintained and made accessible and others are discarded.

9.4.4 Take Into Account Whether the Case is Most or Least Likely for Alternative Explanations

This consideration applies to the ability to generalize the findings of a program evaluation to other contexts in which the program might be instituted. When a program succeeds in its least hospitable conditions, this can provide a warrant for arguing that it is likely to succeed in a wide range of conditions. When it fails in its most favorable context, this suggests a program is unlikely to succeed anywhere. For additional discussion, see Chapter 4.

9.4.5 Make a Justifiable Decision on When to Start

An obvious point in time at which to start an evaluation or establish a baseline is often at the initial implementation of a program. Different parts of a program may have started at different times, however, or they may have started at different times in different regions or for different groups of stakeholders. There can also be time lags between the proposal, approval, and implementation of a program, and during each period stakeholders might start to change their behavior in ways that either enhance or undermine program performance. For example, actors might try to corner the local market and increase the prices of local goods, properties, or services that will be in greater demand once a program starts. In addition, stakeholders may have had incentives to boost or depress some of a program's indicators or measures to try to get initial baseline measures that suit their purposes. When such anticipatory behaviors are possible, it makes sense to consider beginning the evaluation period at the first point in time when actors became aware of the program (which might include private leaks of information, and rumors and misinformation, even before a program is publicly announced).

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

9.4.6 Be Relentless in Gathering Diverse and Relevant Evidence, but Make a Justifiable Decision on When to Stop

The Bayesian logic outlined at the end of this chapter gives rationales for why diverse evidence is important and for deciding on when it is reasonable to stop gathering additional evidence. Essentially, when we assess a particular kind of evidence, each successive piece of this evidence has less potential to strongly change our confidence in different explanations of a case. We will have already updated our views based on the earlier pieces of the same kind of evidence, so each new piece of this kind of evidence is less likely to surprise us, and at some point our time would be better spent looking at a different kind of evidence or a different observable implication of a potential explanation.

At the same time, the appropriate "stopping rule" for looking at a particular kind of evidence depends not just on whether each successive piece of evidence is consistent with the story told by each previous piece, but also on how unexpected that story is in the first place. As the philosopher David Hume wrote, "No testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous than the fact which it endeavors to establish" (Hume 1748, chp. 10).[5] We would thus demand more voluminous, consistent, and diverse evidence to be convinced that a program had an astonishingly strong or weak effect than to be convinced that it does not.

A third consideration for determining a stopping rule for policy-relevant process tracing concerns the question of what is at stake. The higher the consequences of a type I (false positive) or type II (false negative) inference on whether the program worked, the higher the degree of confidence we will seek to establish based on the evidence. It makes sense, for example, to demand more conclusive evidence for medical treatments where lives are at stake than for programs that might at best modestly improve incomes or at worst leave them unchanged.

9.4.7 Combine Process Tracing with Case Comparisons when Useful for the Research Goal and Feasible

Process tracing is a within-case form of analysis, but it can be combined with cross-case comparisons to strengthen inferences. In a "most-similar" case

[5] The astronomer Carl Sagan popularized a pithier formulation: "extraordinary claims require extraordinary evidence."

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

comparison, for example, a researcher selects two cases that are, ideally, similar in the values of all but one independent variable and that have different outcomes on the dependent variable. Before–after comparisons, which compare a preprogram baseline to postprogram outcomes, can be most-similar comparisons if important nonprogram variables do not change in the same time period. The goal in most-similar comparisons is to make an inference on whether the difference on the independent variable – or, here, the program intervention – accounts for the difference on the dependent variable. The key limitation of this design is that even if all but one of the independent variables are closely matched, there may be other untheorized differences between the two cases, including exogenous variables that change in the time period between the inception and the evaluation of a program, that might account for the difference in their outcomes. It is thus important to do process tracing on the independent variable that differs, or the program intervention, to show that it created a causal chain leading up to the outcome. The researcher should also process trace the hypothesized effects of any other potential independent variables that differ between the comparison cases, and to the extent that this reveals that they can be ruled out as causes of the cases' differing outcomes, we can be more confident that the program's theory of change generated the outcome.[6]

9.4.8 Be Open to Inductive Insights

Because the omission of a viable candidate explanation can undermine inferences about a case, it is important to watch for potentially causal variables that were omitted from the initial list of candidate explanations. The feeling of surprise at discovering an unexpected potential causal factor is something to be savored rather than feared, as it signals that there may be something new to be learned about the process that led to the outcome. Cases where the outcome was surprisingly good or unexpectedly poor, or "deviant" or "outlier" cases, are good candidates for process tracing that puts added emphasis on inductive soaking and poking to identify and assess variables whose omission from researchers' or practitioners' prior theories might explain why one or both communities were surprised by the outcome.

[6] Similarly, researchers can use process tracing on "least-similar cases" comparisons, or comparisons among cases with similar measures on only one independent variable and similar outcomes. Here, the researcher can process trace from the common independent variable to the common outcome, and also process trace on any other potential independent variables that are similar to see if they might also account for the outcome.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

9.4.9 Use Deduction to Ask "If the Explanation Is True, What Will Be the Specific Process Leading to the Outcome?"

Researchers need to think concretely about specific hypothesized processes in order to do process tracing well. Social science theories are usually stated in general terms, and it is necessary to adapt them to the case and circumstances at hand and ask what specific sequences and events they would predict if they were to constitute an adequate explanation for the outcome.

Consider the example of microfinance. On one level, the hypothesized mechanism through which such loans work (if they do) is simple: microloans give credit to businesses too small or informal to have access to conventional loans. Yet depending on the details of the microfinance program, several different mechanisms may be at work. In the process of applying for a microloan, applicants might receive feedback that improves their business plans, and those that receive loans may receive further monitoring and advice. Being accepted as a loan recipient might be seen as an indicator of the quality of the applicant's business plan, opening the door to additional credit, whether from social networks or formal financial institutions. If the savings that provide the funds for loans come from local actors who also decide on which loans to make, as in solidarity lending, this can create social pressures – and social resources – for the business to succeed and for loan repayment. Transactions costs, interest rates, inflation, macroeconomic trends, and other factors can affect whether and how microloans work as well. It is necessary to specify concretely how each of these possible mechanisms might have worked in the case at hand, and to outline the observable implications for each, in order to carry out process tracing.

Educational programs provide another example of the importance of thinking concretely about how projects actually work. University scholarship programs aim to provide opportunities for students who could not otherwise afford higher education. It is relatively easy to measure inputs (how many scholarships were given out) and outputs (how many scholarship recipients graduated), but the challenge is to assess how such a program actually works and what its actual effects are compared to the counterfactual world in which the program did not exist. On what basis does it select students for funding? How does it establish and verify the criterion of financial need? Does it also advise students on how to apply to universities and how to prepare for and succeed once they begin attending? Does it get students into programs they would not otherwise attend, or to which they would not even apply without the possibility of a scholarship? What programs were students contemplating

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

or applying to before and after they heard of the scholarship? Might the same students have received scholarships or loans that would allow them to get a university education at the same institutions? Does the scholarship lead to a higher rate of program completion for funded students compared to students who nearly won funding? Were funds provided in a timely way in each semester, or did delay cause dropouts or registration difficulties? Did scholarship students expand the capacities of universities and the numbers of students they accepted, or merely take the place of other students who then had to go to other universities? Did accepting the scholarship open up other funds or resources that the student would have used, creating opportunities for yet other students (including siblings, cousins, etc.)? Such concrete questions get us closer to assessing the actual outcomes that arose and the ways in which they came about.

9.4.10 Remember That Conclusive Process Tracing Is Good, but Not All Good Process Tracing Is Conclusive

When the evidence from a case sharply discriminates among alternative hypotheses – that is, when it is likely to be true under one hypothesis but very unlikely under the alternatives – this allows strong claims that the one hypothesis consistent with the evidence is a strong explanation of the outcome in the case. The evidence is not always strongly conclusive, however, and it is important not to overstate the certainty that the evidence allows. The evidence may be weak or mixed, and it is important to convey how strong the evidence is and how strong the inferences are that the evidence allows. As discussed later in this chapter, this can be expressed in informal terms, such as "smoking gun" versus "straw in the wind" evidence, and "high confidence" or "likely" explanations, or it can be conveyed in numerical point or range estimates of probabilities ranging from zero to one.

In addition, often a combination of factors rather than one factor alone explains the outcome of a case, and it can be difficult to figure out process tracing tests that discriminate among all the possible interactions of the variables of interest. For example, in a particular case of microfinance, it may be that expanded credit alone was sufficient for the outcome, or it may be that this together with business advice from the lender generated the outcome. To distinguish among these, an evaluator would have to think of observable implications that would be consistent with the "credit alone" explanation but not the "both together" explanation, and vice versa.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

A third reason to be careful to not overstate the certitude that the evidence allows is that it is always possible that the outcome is due to an explanation that the evaluator did not consider. As discussed later in this chapter, the Bayesian logic in which process tracing is rooted requires exhaustive and mutually exclusive explanations in order to function completely, and it is never possible to know with certitude that one has considered all the possible explanations. This is one reason that Bayesians do not allow for 100 percent certitude in any inferences.

9.5 Program Evaluation Process Tracing versus Social Science Process Tracing

There is one key difference between program evaluation process tracing and social science process tracing, and it generates both advantages and challenges for program evaluators. This is the fact that the experts who design policy interventions have the opportunity to outline in advance diagnostic indicators that will later provide evidence on whether a program is working as its theory of change suggests. Moreover, officials can require that program implementers begin gathering and reporting evidence on these indicators from the inception of the program or even the preprogram baseline. If the indicators are well designed, and if they also include data on alternative causal processes that might affect program outcomes, this greatly eases the task of program evaluation. Social scientists, in contrast, usually have to devise their own process tracing tests and gather the relevant evidence themselves after the events under study have already taken place.

Predesignation of program indicators can present challenges as well, however. First, indicators may be poorly designed and fail to provide strong evidence on the mechanisms through which the theory of change is expected to operate. Program outcomes can be difficult to conceptualize and measure, which can create a tendency to rely on measuring inputs or outputs instead of outcomes (Castro 2011; Markiewicz and Patrick 2016; Van der Knaap 2016). Diagnostic process tracing evidence is not the same as measures of outputs or outcomes, as it focuses on hypothesized causal mechanisms and processes, but it can overlap with output measures. There can also be a temptation to focus on diagnostic measures that are easy to measure rather than those that provide strong evidence for causal inference.

Second, there is a risk that program managers and other stakeholders will “game” the measurement and reporting of indicators to slant them toward their desired evaluation results. It can be difficult to devise diagnostic measures that provide strong evidence on the causes of program outcomes and that are not also susceptible to gaming. Essentially, this requires devising diagnostic measures that program implementers cannot achieve unless they actually are faithfully carrying out the program in accordance with its theory of change. This can lead to another problem, however: if diagnostic measures are too demanding and detailed, or if program implementers think (rightly or wrongly) that the theory of change is imperfect and that their experience and skills (or changed circumstances) give them better ideas on how to achieve the program's goals, these program managers will face unpleasant choices between following micromanaging guidelines that they think are inappropriate or departing from the prescribed practices and measures. This raises the classic dilemmas concerning how much authority and flexibility to delegate in principal–agent relations, how to monitor agents through management information systems, and whether and how to allow for changes in the middle of program implementation (Honig 2018). While there is no perfect solution to these dilemmas, consulting stakeholders and program managers on the design of appropriate diagnostic measures and putting in place procedures and decision-making processes for modification or adaptation of these measures can minimize the trade-offs between too much and too little delegation and oversight (Gooding et al. 2018).

Perhaps a more common challenge, however, arises when program designers had an under-specified theory of change or gave insufficient attention to developing and gathering evidence on indicators that would make later process tracing and program evaluation easy. Even when a theory of change is well specified, evaluators need to assess its coherence and consider alternative explanations that program managers may not have considered or on which they did not gather evidence. In this regard, program evaluators are often in a position similar to that of social scientists who design and gather evidence on alternative explanations only after the events of interest have taken place.

9.6 Bayesian Logic and Process Tracing

The best practices outlined earlier address the "traditional" process tracing that characterizes almost all published research and completed program evaluations to date. In the last few years, however, methodologists have begun to explore the possibility of applying more explicitly and formally the Bayesian logic that underlies process tracing. There are as yet few applications of this approach to empirical research, and there are strong

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

pragmatic reasons why full formal Bayesian analysis of evidence from case studies is not appropriate in most research settings. Still, it is useful to understand the formal Bayesian logic that informs more informal process tracing practices, as this can lead to better implementation of these less formal practices. In addition, it may be useful to apply more formal Bayesian analysis to a few of the most important pieces of evidence in a study even if it is unduly cumbersome to do so for most of the evidence. While a full discussion of the Bayesian logic of process tracing is beyond the scope of this chapter, the brief outline that follows provides an introduction to the topic.[7]

In Bayesian analysis, probability is conceived of as the degree of belief or confidence that we place in alternative explanations. This is quite different from the standard frequentist statistical conception of probability as representing the likelihood that a sample is or is not representative of a population. In Bayesian analysis of individual case studies, the analyst starts with a "prior," or an initial guess regarding the likelihood that alternative explanations are true regarding the outcome of the case. The analyst uses the logic of the explanations, or of their underlying theories, to estimate how likely particular kinds of evidence are in the possible worlds represented by each explanation. The analyst then uses the laws of conditional probability to translate the likelihood of evidence given alternative explanations into the likelihood of alternative explanations given the evidence. This new, updated estimate of the likelihood that alternative explanations are true is called the "posterior" probability, or simply the posterior.[8]

Bayesianism provides a formal language for discussing the relative strength or probative value of different pieces of evidence. We already have an informal language for this: "smoking gun" evidence strongly supports one

[7] The most complete discussion of Bayesian process tracing to date is Fairfield and Charman (2017). For discussion of Bayesian process tracing in the context of program evaluation, see Befani and Stedman-Bryce (2017).

[8] Using the symbols of probability theory, this paragraph relates to the following version of Bayes Theorem:

$$Pr(\mathrm{P}|\mathrm{k}) = pr(\mathrm{P})pr(\mathrm{k}|\mathrm{P}) \qquad pr(\mathrm{P})pr(\mathrm{k}|\mathrm{P}) + pr(\sim\mathrm{P})pr(\mathrm{k}|\sim\mathrm{P})$$

Notation:
Pr (P|k) is the posterior or updated probability of proposition P given (or conditional on) evidence k
pr(P) is the prior probability that proposition P is true
pr(k|P) is the likelihood of evidence k if P is true (or conditional on P)
pr(~P) is the prior probability that proposition P is false
pr(k|~P) is the likelihood of evidence k if proposition P is false (or conditional on ~P)

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

explanation, but the absence of such evidence does not necessarily reduce confidence in that explanation. Passing a "hoop test" is asymmetrical in the other direction: an explanation is strongly undermined if it fails a hoop test, but we do not necessarily greatly increase our confidence in an explanation that passes a hoop test. These informal examples are points on a continuum: the "likelihood" of evidence taking on a certain value if a theory or an explanation is true can range from 0 to 1, and when we compare the likelihood of evidence under one explanation to its likelihood under an alternative – that is, when we divide the likelihoods – this ranges from 0 to infinity. The more likely evidence is under one explanation, and the less likely it is under the alternatives, the more strongly the discovery of that evidence affirms the one explanation it fits. It is the *relative* likelihood of the evidence under the alternative explanations, or the "likelihood ratio," that matters, not the absolute likelihood that the evidence or data will take on a certain value if one explanation is true.[9]

Bayesian inference, however, is only as good as the information that informs the analysis, which raises the obvious question: How do we estimate the priors and likelihoods? The prior, or our initial guess on the likelihood that a particular theory correctly explains the outcome of a case, in principle represents all of our "background knowledge," or all of our conclusions and intuitions from previous research and experience. In some situations, such as when we have mountains of data, we can use well-informed priors, just as life insurance companies do when they use the ample data at their disposal to estimate life expectancies given a person's age, health habits, and health indicators. Most of the time in social science settings, however, we lack a strong evidentiary basis for estimating priors. One option here is to use uninformed priors – that is, to give each alternative explanation an equal prior (such as a prior of 1/3 if there are three candidate explanations). Another option is to try the analysis with different priors to see how sensitive the conclusions are to the choice of the prior; if the evidence is strong, the estimate of the prior

[9] This relates to the "odds form" of Bayes Theorem, which is mathematically equivalent to the version in the previous footnote but in some ways is more intuitive and easier to work with:

Posterior = Likelihood. Prior Odds
Odds Ratio Ratio Ratio
Or, in the notation of probability:
$Pr(\mathrm{P}|\mathrm{k}) = pr(\mathrm{k}|\mathrm{P}).\ pr(\mathrm{P})$
$Pr(\sim\mathrm{P}|\mathrm{k})\ pr(\mathrm{k}|\sim\mathrm{P})\ pr(\sim\mathrm{P})$

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

might not matter much to the estimate of the posterior (Bayesians call this the "washing out of priors"). A third approach that case study methodologists are beginning to assess is to "crowd source" estimates of priors, whether among subject matter experts or nonexperts.

Estimating likelihoods of evidence is challenging as well. This requires "inhabiting the world" of each hypothesis – that is, assuming that the hypothesis is true and then assessing the likelihood of a piece of evidence given the truth of the hypothesis. Estimating likelihood ratios requires performing this task for multiple hypotheses. On the other hand, it can be easier to assess the relative likelihood of evidence – to ask which of two hypotheses makes the evidence more likely, and even to estimate the ratio of these likelihoods – than to estimate the absolute likelihood of evidence given each hypothesis. As with estimating priors, researchers can try crowd-sourcing estimates of likelihoods.

A third challenge is arranging the alternative explanations, as Bayesian inference requires, in such ways that they are mutually exclusive and exhaustive. This includes explanations that combine several interacting theoretical variables or causal mechanisms, such as agents, institutions, norms, etc. In principle, this is possible for any group of hypotheses. To take a simple example, a criminal investigator might divide the explanations for a murder into four possibilities: the murder could have been committed by suspect A alone, by suspect B alone, by both A and B colluding together, or by neither A nor B. The next step is a bit more complex: the investigator has to think of the likelihood of different pieces of evidence under all these possible explanations, and, ideally, to find evidence that strongly discriminates among the explanations. This can be difficult for murder investigations: the detective has to ask what evidence would point to collusion that would not also be consistent with A or B acting alone. It is arguably even more challenging for social science researchers who are evaluating various combinations of structural, normative, macroeconomic, managerial, and other factors that can contribute to the success or failure of development programs.

A final difficulty with formal Bayesian analysis is that the calculations it requires become tedious and lengthy to write up and to read even for a small number of pieces of evidence and alternative explanations, and much more so for multiple pieces of evidence and explanations. For this reason, even the methodologists who have begun to explore formal Bayesian process tracing argue against trying to implement it fully for all the evidence (Fairfield and Charman 2017).

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

Still, it can be useful to do formal Bayesian analysis on one or a few of the pieces of evidence that a researcher judges to be most powerful in discriminating among alternative explanations, as this can make the analysis more transparent. Specifically, understanding the Bayesian logic of process tracing can contribute to better process tracing practices in at least four ways. First, Bayesian logic provides a clear philosophical warrant for much of the practical advice methodologists have given regarding traditional process tracing, including the ten best practices of traditional process tracing discussed earlier. One reason to initially consider a wide range of alternative explanations, for example, is that failing to consider a viable explanation can bias the estimates of the likelihoods, and thus the posterior estimates, of all the explanations the analyst does consider. Bayesianism also gives a clear explication of what constitutes strong evidence, of why diverse and independent evidence is important, of the trade-offs involved in stopping too soon or too late in gathering and analyzing evidence, and of why we should never be 100 percent confident in any explanation.

Second, Bayesianism leads to counterintuitive insights. Evidence that is consistent with an explanation, for example, can actually make that explanation less likely to be true if the same evidence is even more consistent with an alternative explanation. Also, numerous pieces of weak evidence (or what might be called "circumstantial evidence" in a court), if they all or mostly point in the same direction, can jointly constitute strong evidence that considerably changes our confidence in alternative explanations.

Third, formal Bayesian analysis, even if it is done only on a few key pieces of evidence, provides a transparent form of inference that allows researchers and their readers or critics to identify exactly why their inferences diverge when they disagree on how to update their confidence in explanations in light of the evidence. Researchers and their readers can disagree about their priors, the likelihood of evidence under alternative explanations, and the interpretation or measurement of the evidence itself. Leaving estimates and interpretation of each of these ambiguous obscures where authors and readers agree and disagree. Making judgments on each of these clear, in contrast, can prompt researchers and their critics to reveal the background information that underlies their judgments, which can narrow areas of disagreement.

The fourth, and perhaps strongest, rationale for learning Bayesian analysis is that it illuminates the logic that traditional process tracers have used informally all along in order to make causal inferences form individual cases, and it can help them to use it better. Research on the psychology of decision-making indicates that people often make mistakes when they try to be intuitive

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

Bayesians or first attempt formal Bayesian analysis (Casscells, Schoenberger, and Grayboys 1978). Other research shows that deeper training in Bayesian analysis can help improve forecasting (Tetlock and Gardner 2015). Additional research indicates that a few simple practices consistent with Bayesian process tracing, such as actively considering alternative explanations, can help debias judgments (Hirt and Markman 1995).

9.7 Conclusion

Process tracing and program evaluation, especially forms of evaluation that emphasize contribution analysis, have much in common. Both involve inferences on alternative explanations of outcomes of cases. It is not accidental that the evaluation community has taken a growing interest in process tracing, or that process tracing methodologists have become interested in program evaluation. The best practices developed in traditional social science process tracing are applicable, with modest adaptations, to the task of program evaluation. The biggest difference is that in contrast to researchers doing process tracing in the social sciences, program evaluators may have the opportunity to designate in advance, and to require reporting upon, diagnostic indicators about alternative processes as well as measures of inputs, outputs, and outcomes. This can make later evaluation easier, but it can also introduce potential distortions and biases as program managers and stakeholders might “game the system” once they know what measures will be tracked. Program designers and evaluators need to be creative and flexible in designing indicators that are useful in subsequent program evaluations, that cannot be achieved without also achieving the desired results at which a program aims, and that do not become a straightjacket on program managers when modifications to a program can better achieve its goals.

Program evaluators can benefit as well from exploring the emerging literature on formal Bayesian process tracing. This literature clarifies the logic behind traditional process tracing methods, and it is beginning to explore and outline new practices, such as crowd-sourcing of estimates of priors and likelihood ratios, that might further strengthen process tracing. Although formally analyzing the weight of every piece of evidence is impractical, it can be useful to formally assess a few of the strongest pieces of evidence. This can contribute to more logically consistent and analytically transparent assessments of alternative explanations of program outcomes.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

References

Astbury, B. and Leeuw, F. L. (2010) "Unpacking black boxes: Mechanisms and theory building in evaluation," *American Journal of Evaluation*, 31(3), 363–381.

Bamanyaki, P. A. and Holvoet, N. (2016) "Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives," *Evaluation*, 22(1), 72–90.

Barnett, C. and Munslow, T. (2014) "Process tracing: The potential and pitfalls for impact evaluation in international development," Summary of a workshop held on May 7, 2014, IDS Evidence Report 102, Brighton: IDS.

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

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

Bennett, A. (2013) "The mother of all isms: Causal mechanisms and structured pluralism in international relations theory," *European Journal of International Relations*, 19(3), 459–481.

Bennett, A. and Checkel, J. (2015) "Process tracing: From philosophical roots to best practices" in Bennett, A. and Checkel, J. (eds.) *Process tracing: From metaphor to analytic tool.* Cambridge: Cambridge University Press, pp. 3–37.

Busetti, S. and Dente, B. (2017) "Using process tracing to evaluate unique events: The case of EXPO Milano 2015," *Evaluation*, 23(3), 256–273.

Casscells, W., Schoenberger, A., and Grayboys, T. B. (1978) "Interpretation by physicians of clinical laboratory results," *The New England Journal of Medicine*, 299(1978), 999–1001.

Castro, M. F. (2011) "Defining and using performance indicators and targets in Government M and E systems." Washington, DC: The World Bank. PREM Notes and Special series on the Nuts and Bolts of Government M&E Systems; No. 12.

Cook, T. D. (2018) "Twenty-six assumptions that have to be met if single random assignment experiments are to warrant 'gold standard' status: A commentary on Deaton and Cartwright," *Social Science & Medicine*, 210(2018), 37–40.

Dalkin, S. M., Greenhalgh, J., Jones, D. et al. (2015) "What's in a mechanism? Development of a key concept in realist evaluation," *Implementation Science*, 10(49), 1–7.

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

Dunning, T. (2015) "Improving process tracing: The case of multi-method research" in Bennett, A. and Checkel, J. (eds.) *Process tracing: From metaphor to analytic tool.* Cambridge: Cambridge University Press, pp. 211–236.

Fairfield, T. and Charman, A. (2017) "Explicit Bayesian analysis for process tracing: Guidelines, opportunities, and caveats," *Political Analysis*, 25(3), 363–380.

Gooding, K., Makwinja, R., Nyirenda, D. et al. (2018) "Using theories of change to design monitoring and evaluation of community engagement in research: Experiences from a research institute in Malawi," *Wellcome Open Research* 3(8), available at https://wellcomeopenresearch.org/articles/3-8 (accessed November 22, 2019).

Hirt, E. R. and Markman, K. D. (1995) "Multiple explanation: A consider-an-alternative strategy for debiasing judgments," *Journal of Personality and Social Psychology*, 69(6), 1069–1086.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press

Honig, D. (2018) *Navigation by judgment: Why and when top down management of foreign aid doesn't work*. New York: Oxford University Press.

Hume, D. (1748) *An enquiry concerning human understanding*. New York: Oxford University Press. Posthumous edition 1777; edited with introduction by L. A. Selby-Bigg, 1902, reproduced through Project Gutenberg. Accessed at www.gutenberg.org/files/9662/9662-h/9662-h.htm.

Ioannidis, J. P. A. (2005) "Why most published research findings are false," *PLOS Med*, 2(8), e124. Available at https://doi.org/10.1371/journal.pmed.0020124 (accessed November 22, 2019).

Machamer, P., Darden, L., and Craver, C. F. (2000) "Thinking about mechanisms," *Philosophy of Science*, 67(1), 1–25.

Markiewicz, A. and Patrick, I. (2016) *Developing monitoring and evaluation frameworks*. Thousand Oaks, CA: Sage.

Mendoza, A. A. and Woolcock, M. (2014) Integrating qualitative methods into investment climate impact evaluations. World Bank Policy Research Working Paper No. 7145. Available at https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-7145 (accessed November 22, 2019).

Pawson, R. and Tilley, N. (1997) *Realistic evaluation*. Thousand Oaks, CA: Sage.

Punton, M. and Welle, K. (2015) "Straws-in-the-wind, hoops and smoking guns: What can process tracing offer to impact evaluation?" Centre for Development Impact Practice Paper 10, Brighton: IDS, 1–8.

Salmon, W. (1998) "Scientific explanation: Causation and unification" in Salmon, W. (ed.) *Causality and explanation*. New York: Oxford University Press.

Salmon, W. (2006) *Four decades of scientific explanation*. Pittsburgh: University of Pittsburgh Press.

Schmitt, J. and Beach, D. (2015) "The contribution of process tracing to theory-based evaluations of complex aid instruments," *Evaluation*, 21(4), 429–447.

Shadish, W., Cook, T., and Campbell, D. T. (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Stern, E., Stame, N., Mayne, J. et al. (2012) Broadening the range of designs and methods for impact evaluations. London: DFID Working Paper 38.

Tetlock, P. and Gardner, D. (2015) *Superforecasting: The art and science of prediction*. New York: Broadway Books.

Van der Knaap, P. (2016) "Responsive evaluation and performance management: Overcoming the downsides of policy objectives and performance indicators," *Evaluation*, 12(3), 278–293.

Waldner, D. (2012) "Process tracing and causal mechanisms" in Kincaid, H. (ed.) *The Oxford handbook of philosophy of social science*. New York: Oxford University Press.

Wauters, B. and Beach, D. (2018) "Process tracing and congruence analysis to support theory-based impact evaluation," *Evaluation*, 24(3), 284–305.

https://doi.org/10.1017/9781108688253.010 Published online by Cambridge University Press