CHAPTER 6

Developing Empirical Tests of Causal Mechanisms

Chapter 4 dealt with how we should develop theories, where a causal theory (X → Y) should be reconceptualized into a theorized causal mechanism composed of a set of parts, each of which can be thought of as a hypothesis (h) about which we have expectations of the prior probability of its existence (p(h)). In the Bayesian inferential logic described in chapter 5, the purpose of empirical tests is to update our degree of confidence in a hypothesis in light of the empirical evidence that has been found. Our ability to update the posterior probability of a hypothesis is contingent on the probability of evidence and the ability of our empirical tests to discriminate between evidence that supports h and alternative hypotheses (~h). This chapter discusses how empirical tests of hypothesized mechanisms are developed in process-tracing research.

Testing theorized causal mechanisms involves formulating case-specific predictions about the expected observable manifestations of each part of a causal mechanism that we should see if it is present. We define these observable manifestations as evidence that we should expect to find in the case if each part of a causal mechanism is present. In developing these case-specific predictions, we deploy our contextual knowledge of individual cases. Predicted evidence is analogous to what Adcock and Collier (2001) term *empirical indicators* for theoretical concepts, although we use the term to refer to the predicted evidence for each part of a mechanism instead of just X and Y.

We begin this chapter by further developing the Bayesian foundations for empirical tests of theorized causal mechanisms in process-tracing, illustrating how our ability to update our confidence in the presence of a causal mechanism and all of the parts depends on (1) the probability of the evi-

dence (p(e)), (2) the likelihood ratio (p(e|~h)/p(e|h)), and (3) the theoretical prior. The second section discusses the different types of evidence utilized in process-tracing analysis, distinguishing among pattern, sequence, trace, and account evidence. The third section develops four test types for developing predictions for what evidence we should expect to find if h is valid, arguing that we should attempt to maximize both the certainty and uniqueness of our predictions to maximize our ability to update our confidence in the hypotheses in light of empirical evidence (van Evera 1997). We provide practical suggestions for maximizing both the certainty and the uniqueness of our predictions. Finally, we illustrate the challenges relating to developing empirical tests using a bureaucratic politics mechanism as an example.

6.1. Bayesian Updating

The key to testing theories in process-tracing analysis is to maximize the inferential power of our empirical tests for whether the parts of a hypothesized causal mechanism exist. The stronger the test, the more we can update our degree of confidence in the presence/absence of the parts of the hypothesized mechanism. To understand test strength and in particular the power of evidence to update our confidence, we need to look more closely at three terms in Bayesian theory: the probability of evidence, the likelihood ratio relating hypotheses to evidence, and theoretical priors.

The importance of the probability of evidence (p(e)) is most clearly seen in Bayes's original theorem, or what has been termed the first form of the Bayesian theorem (Howson and Urbach 2006: 20–21). This is a simpler version of the theorem than the one presented in chapter 5.

$$p(h|e) = \frac{p(e|h)\ p(h)}{p(e)}$$

Here the posterior (p(h|e)) is equal to the probability of the evidence conditional on the hypothesis being valid multiplied by the probability of the hypothesis, divided by the probability of the evidence taken by itself. Because p(e) is in the denominator, as the probability of the evidence decreases, the ability of evidence to update the posterior increases, other things equal. More surprising evidence (low p(e)), if found, results in larger increases in our confidence in a hypothesis than less surprising evidence. Given its unlikelihood, a man-bites-dog story, if found, has a stronger inferential weight

than a more typical dog-bites-man story. We return to the question of the probability of evidence p(e) in our discussion of evaluating evidence in chapter 7. The second term of interest in Bayesian updating is the likelihood ratio, the effect of which is most clearly seen in the third form of Bayes's theorem from chapter 5. The likelihood ratio is p(e|~h)/p(e|h). The likelihood ratio should be read as the probability of finding the predicted evidence (e) if the alternative hypothesis (~h) is true in comparison to the probability of finding the evidence if the hypothesis is true. The ratio captures an empirical test's ability to discriminate between predicted evidence that supports h and ~h. If h predicts e, the occurrence of e will raise our confidence in the validity of h, depending on the size of likelihood ratio. When p(e|h) is high and p(e|~h) is low, finding e results in a large increase in confidence. While the ratio depicts ~h as a single alternative, it can also be defined as any plausible alternative to h.

$$p(h|e) = \frac{p(h)}{p(h) + \frac{p(e|\sim h)}{p(e|h)}p(\sim h)}$$

In practical research situations, we often lack a clearly demarked alternative hypothesized explanation of a part of a causal mechanism, and even if we have a relatively clear alternative, it is often difficult to design tests that do not privilege h over ~h, or vice versa. For example, in testing ideational versus material interest mechanisms as explanations for the lack of force used by the Soviets in 1989 to stop the revolutions in the Warsaw Pact countries, Tannenwald suggests that tests of mechanisms that take material factors as their starting point privilege materialist explanations, and vice versa (2005: 22–23). She states that "the larger issue of whether one starts with ideas or material factors in assessing the role of ideas is ultimately impossible to resolve and will remain a fundamental point of disagreement among scholars." However, she pragmatically attempts nonetheless to develop balanced tests of h and ~h—for example, by looking at the range of options considered by actors to see whether certain options were considered unthinkable (role of ideas) or whether all plausible options were considered (material interests) (23, 24).

Even if there is a well-developed alternative, the testing of two contending hypotheses for a part of a mechanism (h and ~h) against each other can degenerate into the sort of gladiatorial contest seen in the movie *Mad Max beyond Thunderdome,* where two men enter, one man leaves. In these

gladiator-like contests, ~h is often depicted in such a highly simplified form that its fate is sealed even before the analysis starts (Checkel 2012).

Therefore, we agree with Bennett's suggestion that we should cast our net widely for alternative plausible explanations of predicted evidence of a particular part of a mechanism, attempting to avoid crude either/or analyses (2008a: 707). The questions that we can ask when testing each part of a mechanism include, Can plausible alternative mechanisms be formulated? Are there any other plausible explanations for the predicted evidence?

The final step is to define theoretical priors, understood as the expected probability that a given hypothesized mechanism (and its parts) is valid (p(h)). When we are engaging in causal inferences in the empirical analysis, the process of updating our level of confidence in the validity of the parts of a mechanism is informed by our prior beliefs about our confidence in the presence of a mechanism as a whole.[1]

If we are testing whether a democratic peace mechanism exists, our prior would be our confidence, based on what we know, in the existence of a mechanism. But answers to the question of the level of prior expectations vary across different research traditions in political science. Regarding democratic peace, skeptics from the realist tradition of international relations contend that the thesis has a very low prior probability of being true (low p(h)) despite the large number of quantitative and qualitative studies that have been undertaken. For example, realists have contended that the found correlation between democracy (X) and peace (Y) is the product of confounding variables such as the distribution of democracies on one side of the divide during the Cold War (Farber and Gowa 1997; see also Layne 1994; Rosato 2003). A realist would therefore start a process-tracing analysis with a low prior as a starting point in a particular case.

As discussed briefly in chapter 5, the use of priors introduces a degree of unavoidable subjectivity into the analysis. In Bayesian statistical analysis, the standard procedure is to test posteriors' sensitivity on prior estimates by simulating how the posterior changes depending on different priors (Jackman 2004). We cannot do this in qualitative research; therefore, when the literature exhibits significant disagreement about a prior, the best bet is to use conservative estimates of p(h) relative to p(~h). In other words, we prejudice our analysis away from h and toward alternative theories (Lynch 2005).

Further, priors inform all scientific research, and by explicitly stating them, we make their incorporation transparent (Chalmers 1999; Howson and Urbach 2006). For example, in process-tracing analysis, we should state that "based upon prior research, we are a priori relatively confident that theoretical proposition X is valid." This prior then informs the subsequent

evaluation of the degree of updating of our confidence in the theory after it has been empirically tested in our case study.

Most important, priors tend to wash out after repeated meetings with empirical evidence (Howson and Urbach 2006: 298). A realist and liberal scholar who engaged in research on the democratic peace mechanism would start with different priors but after repeated empirical studies would converge on the same posterior, assuming that strong empirical tests were developed that did not privilege one of the theorized mechanisms and that the empirical evidence was accurate.

6.2. Types of Evidence in Process-Tracing

Testing whether causal mechanisms are present in process-tracing analysis involves investigating whether our theory-based predictions for what we should see in the empirical record are matched in reality. Do we find the predicted evidence of the parts of the mechanism? Here, the term *relevant evidence* used in evidence law is appropriate. U.S. Federal Rule of Evidence 401 defines *relevant evidence* as "any evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without evidence."

In comparison to data-set observations, predicted evidence in process-tracing is more analogous to what Collier, Brady, and Seawright (2010b: 184–88) term "causal-process observations" or what Bennett refers to as process-tracing observations (2006: 341). Pieces of evidence in process-tracing are noncomparable, as they take many different forms depending on what type of evidence is best suited to testing a particular hypothesized part of a causal mechanism. As discussed earlier, observations only become evidence after they have been evaluated using our contextual knowledge of the specific case ($o + k \rightarrow e$).

Four distinguishable types of evidence are relevant in process-tracing analysis: pattern, sequence, trace, and account. Pattern evidence relates to predictions of statistical patterns in the evidence—for example, in testing a mechanism of racial discrimination in a case dealing with employment, statistical patterns of employment would be relevant for testing parts of the mechanism. Sequence evidence deals with the temporal and spatial chronology of events predicted by a hypothesized causal mechanism. For example, a test of a hypothesis could involve expectations of the timing of events where we might predict that if h is valid, we should see that that event b took place after event a. However, if we then found that event b took place before event

a, the test would suggest that our confidence in the validity of this part of the mechanism should be reduced (disconfirmation).

Trace evidence is evidence whose mere existence provides proof that a part of a hypothesized mechanism exists. For example, the existence of the official minutes of a meeting, if authentic, provides strong proof that a meeting took place. Finally, account evidence deals with the content of empirical material, such as meeting minutes that detail what was discussed or an oral account of what took place in a meeting. Khong (1992) uses this type of evidence when he illustrates the range of policy options considered by U.S. decision makers in the escalation of the Vietnam War.

In designing empirical tests, we need to state clearly what types of evidence we should expect to see if a hypothesized part of a mechanism exists. In the Moravcsik example (see chapter 4), a test of the hypothesis regarding the Commission possessing privileged information can be operationalized as "We should expect to see that the Commission in the most sensitive areas of negotiations was much better informed about the content and state of play of the negotiations than governments, possessing more detailed issue briefs and more accurate and updated information on the state of play of the negotiations." In a case testing this hypothesis, we should expect to see pattern evidence in the form of a greater number of internal study papers and papers documenting the state of play in the Commission archives than in national governmental archives. Here, the quantity of observations forms the evidence, with contextual knowledge playing only a lesser role in evaluating that each document was a study paper or a description of the state of play (what different actors wanted at specific points in the negotiations). Second, another type of relevant evidence would be account evidence, where we should expect that participants in the negotiations would verify in interviews that the Commission was better informed both about the content of issues and the state of play of the negotiations. Contextual knowledge would, for example, be used to evaluate each interview to detect potential sources of bias.

6.3. Test Strength

The logic of empirical testing in process-tracing is that if we expected X to cause Y, each part of the mechanism between X and Y should leave the predicted empirical manifestations that can be observed in the empirical material. Detecting these manifestations, or fingerprints, requires the development of carefully formulated case-specific predictions of what evidence we should expect to see if the hypothesized part of the mechanism exists. The

key difference between predicted evidence as used in process-tracing and other in-depth case study methods such as the congruence method is that in process-tracing, predictions should be formulated in a manner that captures both the entity and the activity involved in each part of a causal mechanism, whereas congruence tests typically test the same prediction for X and Y by themselves at different times during an empirical process (t_0, t_1, . . . t_n).

In other words, empirical tests in process-tracing should be designed in a manner that captures traces of the transmission of causal forces through the theorized causal mechanism. Bayesian logic suggests that when we design empirical tests, we need to maximize the inferential power of evidence to update our confidence in the validity of a hypothesis (Good 1991). This involves making strong predictions about (1) what evidence we should expect to see if a part of a causal mechanism exists and (2) what counts as evidence for alternative hypotheses, while taking into consideration (3) what we can conclude when the predicted evidence is not found—that is, what counts as negative evidence (~e). In operationalizing the parts of a causal mechanism into a series of tests, we need to think about our ability to use the empirical tests to update our confidence in the validity of h when we find e and when e is not found (~e).

Van Evera has introduced a helpful terminology for evaluating test strength that can be adapted for use in process-tracing testing and is compatible with the underlying logic of Bayesian updating (1997: 31–34).

First, what types of inferences can be made when we find the predicted evidence? Does finding e increase our confidence in the existence of the part of the mechanism in relation to plausible alternatives? Van Evera uses the term *unique predictions* to refer to the formulation of empirical predictions that do not overlap with those of other theories. Uniqueness corresponds to the likelihood ratio, where predictions are developed that maximize the value of p(e|h) in relation to p(e|~h). Therefore, if we have a hypothesis formulated in a manner that is highly unique and if we find the predicted evidence, our confidence increases in the presence of the part of the mechanism. In the Sherlock Holmes example from chapter 5, the prediction of finding that the dog that did not bark was highly unusual, with other explanations for the lack of bark quite implausible (p(e|h) was high relative to p(e|~h)).

The second dimension relates to what types of inferences we can make when we do not find the predicted evidence. If we find ~e, can we update our confidence that the part of the mechanism does not exist? What empirical evidence do we have to see for h to be valid? Van Evera (1997) terms this a *certain prediction,* meaning that the prediction is unequivocal and the prediction (e) must be observed or the theory fails the empirical test. Logi-

cally, when p(e|h) = 1, then p(h|~e) = zero, given that p(h) > 0 (Howson and Urbach 2006: 93). This means that if our prediction is maximally certain and we find ~e, our posterior confidence in h is 0, meaning that we have maximally disconfirmed the existence of h. Popper was so fascinated by this logic that he used it as the basis for his falsification principle. However, falsification is a misguided strategy, given that we never can be 100 percent sure that the evidence collected was accurate or that the test was 100 percent certain (93). Disconfirmation should therefore be understood as a matter of degree. If we find ~e, we downgrade our confidence in the existence of the theorized part of the mechanism in relation to the degree of certainty of the empirical prediction. Yet given the ambiguity of social science data, we can never falsify a theory.

Van Evera classifies different types of tests of predictions along these two dimensions, resulting in four ideal-typical types of tests: straw-in-the-wind, hoop, smoking gun, and doubly decisive (1997: 31–34). These dimensions should actually be understood as continuums, based on the idea that through diligent refining of our predictions, we can push test strength further in the direction of more certain and more unique tests (see figure 6.1). At the same time, 100 percent certain or unique tests are a will-o'-the-wisp, impossible to attain as a result of the impossibility of perfect measurement of social phenomena. Therefore, test strength is a matter of degree and is best represented as a continuum.

The weakest type of test is a straw-in-the-wind test. These are empirical predictions that have a low level of uniqueness and a low level of certainty. These tests do little to update our confidence in a hypothesis irrespective of whether we find e or ~e, as both passed and failed tests are of little if any inferential relevance. Bennett (2008a) suggests an example of a straw-in-the-wind test for whether ideas or material interests mattered for the Soviet nonuse of force in 1989. Looking exclusively at the views of Soviet premier Gorbachev and testing predictions from both hypotheses is inconclusive since "the policy views of any one individual, even an individual as historically important as Gorbachev, cannot definitively show that material incentives rather than ideas were more important in driving and shaping changes in Soviet policies" (716).

Hoop tests involve predictions that are certain but not unique; the failure of such a test (finding ~e) reduces our confidence in the hypothesis, but finding e does not enable inferences to be made (Bennett 2006; Van Evera 1997). For example, in criminal trials, questions such as, "Was the accused in the town on the day of the murder?" and "Was the suspect too big to squeeze through the window through which the murderer entered the house?" are

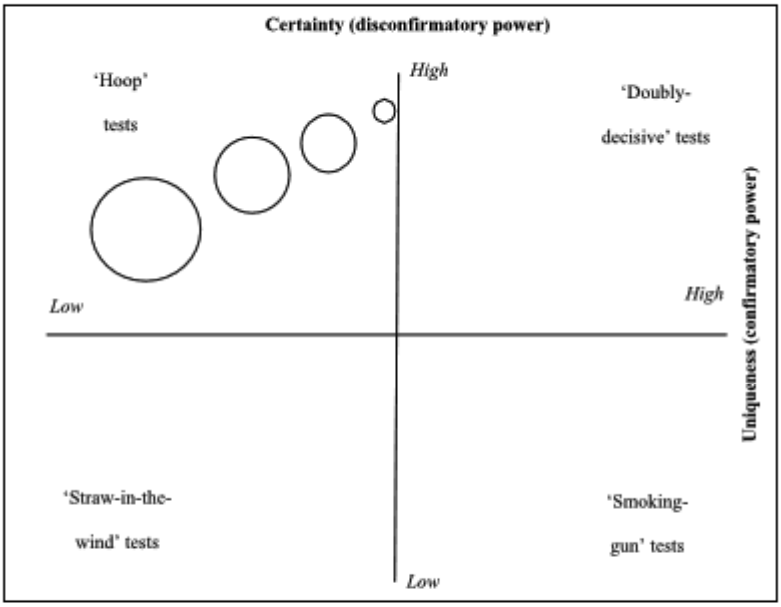

Fig. 6.1. Types of tests of parts of a causal mechanism. (Based on Van Evera 1997: 31–34.)

hoop tests through which a hypothesis would need to pass. However, finding that the suspect was in the town does not confirm the suspect's guilt, whereas finding that the suspect was out of the state would suggest innocence, contingent on the accuracy of the evidence supporting the suspect's alibi. In practice, hoop tests are often used to exclude alternative hypotheses.

Figure 6.1 depicts hoop tests as having different diameters, illustrating that as we strengthen the uniqueness and certainty of our hoop tests, we increase test strength. As the hoop through which the hypothesis must jump becomes progressively smaller, our ability to update our degree of confidence in the validity of the hypothesis is updated when we find either e or ~e. A prediction that a suspect is in the state at the time of a crime is a large hoop, whereas a much tighter hoop would be whether the suspect was in the neighborhood at the time a crime was committed. As the later test is much harder for a hypothesis to jump through, the inferential value of the test is greater.

Smoking gun tests are highly unique but have low or no certainty in their predictions. Here, passage strongly confirms a hypothesis, but failure does

not strongly undermine it. A smoking gun in the suspect's hands right after a murder strongly implicates the suspect, but if we do not find the gun, the suspect is not exonerated. In Bayesian terms, the likelihood ratio is small, with finding e given h highly probable, whereas -h is highly improbable, thereby greatly increasing our confidence in the validity of h if we find e. However, employing this type of test is a high-risk strategy given that a small likelihood ratio usually implies relatively improbable evidence (p(e) low), which by its nature makes e very difficult to find. Yet if we do not find e, we are unable to update our confidence in the validity of h. In other words, the dilemma is that unless we find e, a smoking gun test is pointless (no updating possible), but if we design a weaker test that increases the probability that we will find e (i.e., p(e) higher), our ability to update the posterior if we find e is reduced considerably.

Finally, doubly decisive tests combine both certainty and uniqueness. If the evidence is not found, our confidence in the validity of the hypothesis is reduced; at the same time, the test discriminates strongly between evidence that supports the hypothesis and alternatives (small likelihood ratio). In a criminal trial, a doubly decisive test could be the tape of a high-resolution surveillance camera from the scene of the crime. The prediction that the suspect should be identifiable in the tape is relatively certain, for if the expected evidence (suspect captured on the tape) is not found (and the tape has not been tampered with), we can with reasonable confidence conclude that the suspect is innocent. Further, if the tape shows the suspect committing the crime, this test confirms guilt unless there is evidence suggesting that the suspect was acting under duress (high uniqueness).

Doubly decisive tests are therefore ideal. However, in real-world social science research, it is also almost impossible to formulate predictions in such a manner given the difficulty of finding and gaining access to the type of empirical evidence that would enable doubly decisive tests. Furthermore, inverse relationships often exist between the uniqueness and certainty of tests in that the more unique the empirical predictions, the less likely we are to find the evidence, and vice versa.

We advocate seeking to maximize the levels of both certainty and uniqueness. However, when push comes to shove and we are forced to choose between certainty and uniqueness, we suggest prioritizing certainty over uniqueness in process-tracing test designs based on Bayesian logic. As we argue in chapter 3, each part of a causal mechanism should be seen as individually necessary. While passing a form of smoking gun test would substantially increase our confidence in the presence of a given part of a causal mechanism, the failure to find a smoking gun tells us nothing. Because we

are operating with what are in effect single-case studies for each part of a mechanism, we need to design tests that have a relatively high degree of certainty (hoop tests), since the absence of the evidence (~e) of a certain prediction allows us to infer with a reasonable degree of certainty that the part of the mechanism was not present.

In addition, hoop tests can be employed in clusters in process-tracing research, where multiple independent hoop tests of what we must find in the evidence (certainty) are formulated for each part of a mechanism. When hoop tests are combined to test h, the result is an additive effect that increases our ability to update our confidence in the validity of h given that the probability of a nonvalid hypothesis surviving multiple independent hoop tests falls after each successive hoop. The greater the variety of independent evidence collected, the stronger the support for the hypothesis, other things equal (Howson and Urbach 2006: 125–26).

An Example—Improving Test Strength

We seek to maximize the strength of our tests of each part of a causal mechanism. In chapter 5, we discussed Moravcsik's (1998, 1999) theorization on the influence of supranational actors in EU negotiations. Moravcsik (1999) tests whether a causal mechanism exists that posits that supranational actors such as the Commission gain influence in EU negotiations because of their relative informational advantages. One part of the causal mechanism is conceptualized as "The Commission has privileged access to information."

A hypothetical example of a straw-in-the-wind test that puts forward uncertain and nonunique predictions to test the hypothesized part could be that we should "expect to see that the Commission has many civil servants." However, this is a weak test on both of van Evera's dimensions. First, the test has a low level of certainty, since the information we need is not the number of civil servants per se but whether these officials have relative informational advantages vis-à-vis national civil servants. Second, the uniqueness of the test is low, as finding many civil servants can also be explained by competing theories such as intergovernmentalist theory. Intergovernmentalism would argue that most of the civil servants are translators and therefore do not contribute to granting the Commission privileged access to information.

The test could be improved on both dimensions so that it approaches a doubly decisive test. This test could be formulated as we should "expect to see that the Commission in the most sensitive areas of negotiations was much better informed about the content and state of play of the negotia-

tions than governments, possessing more detailed issue briefs and more accurate and updated information on the state of play." This test has a higher degree of certainty, as finding this evidence is vital in making the argument that the Commission possesses privileged information. Further, the predictions are quite unique in that this phenomenon is something we would not expect to see if competing alternative theories such as intergovernmentalism were correct. In Bayesian terms, formulating the hypothesis in this way help us to increase the likelihood ratio and thereby increase our confidence in the theory if we find evidence that supports our theory; at the same time, the level of certainty allows us to update our confidence in the nonexistence of the part of the mechanism if we do not find e but find ~e.

Conclusions

In operationalizing empirical tests of causal mechanisms, we develop predictions about what we should expect to see in the empirical record if a hypothesized part of a causal mechanism is present (predicted evidence), formulated in a manner that maximizes the level of certainty and uniqueness. When developing predictions of what type (or types) of evidence we should see for each part of the mechanism, it is useful to think in terms analogous to playing both sides (prosecutor and defense) in a criminal trial. What evidence must appear in the empirical record if the hypothesized part of the mechanism is present (certainty)? If found, can we explain the evidence using alternative hypotheses (~h) (uniqueness)?

The testing process takes place for each part of a causal mechanism. As mentioned in chapter 5, the strength of our inferences about the presence of a causal mechanism depend on the ability of the test of each part of a mechanism to update significantly our confidence (up or down) in its presence. Therefore, our inferences about the presence of a causal mechanism are only as strong as the weakest test, making it vital that we strive to maximize test strength for each part. In practice, we often find that testing the parts in between is much more difficult than testing the start and the outcome of a causal mechanism. When faced with this challenge, we suggest suitable caution regarding inferences about the presence and workings of the mechanism. The analysis must clearly flag the weakest link(s)—for example, by stating that part 3 can only be tested using a straw-in-the-wind test that does little to update our confidence in the presence/absence of the part. At the same time, we must be realistic, and if we can provide strong tests (ide-

ally doubly decisive) of many of the parts of a mechanism and the predicted evidence is found, we can cautiously conclude that we have updated our confidence in the presence of the mechanism as a whole.

Testing in explaining-outcome process-tracing is often a much messier task given that we usually do not possess a simple mechanism; instead, we want to test the presence of multiple overlapping mechanisms with different parts, including nonsystematic mechanisms. However, in line with the ontological and epistemological position of pragmatism taken by scholars engaging in explaining-outcome process-tracing (see chapter 2), we also need to acknowledge that empirical tests here are usually much more instrumental. Theorized mechanisms are seen as heuristic tools that guide our inquiry toward a persuasive explanation (Humphreys 2010: 268). Empirical analysis is closer to what can be thought of as calibrating a case-specific explanation in an iterative research design to account for discrepancies between the theorized mechanisms and the data (Jackson 2011: 146–47) (see chapter 2). However, Bayesian logic and the strength of tests still provide the foundations for the causal inferences that are made, but they are used in a looser and more pragmatic fashion than in theory-centric variants of process-tracing, where the aim is the more instrumental one of assessing the similarity between the predicted evidence, what is found, and the ability of a given theorized mechanism to account for those results in terms of the likelihood ratio.

In the rest of the chapter, we present an extended example of how we should conceptualize and operationalize causal mechanisms in theory-testing process-tracing, illustrating the many challenges we face in practical research situations.

6.4. An Extended Example of Conceptualization and Operationalization of Causal Mechanisms: Studying Bureaucratic Politics

In this example, we attempt to build a design that can test the presence/absence of a hypothesized causal mechanism for bureaucratic politics in a single-case process-tracing study. The example illustrates some of the challenges in utilizing process-tracing methods and designing tests that can measure and evaluate the existence of the different parts of a hypothetical causal mechanism. The specific case involves the interministerial coordination of national positions on the issue of the transfer of sovereignty to the EU within EU constitutional negotiations. The basic question is whether

politicians are in control of the formulation of national positions or whether bureaucratic self-interest plays a role in the translation of societal preferences into negotiable national positions.

Conceptualizing a Bureaucratic Politics Mechanism

There is a wealth of literature within public administration research and foreign policy analysis (FPA) that suggests that how the state is organized matters in domestic and foreign policy, with the manner of organization viewed as a crucial intervening variable between inputs (societal preferences) and outcomes (national positions). In the theory of bureaucratic politics, ministries attempt to pursue their parochial institutional interests in a bargaining game of pulling and hauling (Allison and Zelikow 1999; Bendor and Hammond 1992: 317–18; Caldwell 1977: 95; Hermann et al. 2001; Jones 2008; Michaud 2002; Peters 1995; Preston and 't Hart 1999; Rosati 1981). Policy outcomes are not merely the result of a rational matching of national interests and solutions but are better seen as the product of internal conflict, compromise, and negotiation among competing ministries. This definition differs from the revised model developed by Allison and Zelikow (1999) that they term governmental politics, which became more focused on the high-level political battles between different senior political figures (cabinet officials) and the president rather than with the lower-level interministerial battles among bureaucrats within a government. Here we are interested in explaining how interministerial battles (bureaucratic politics) produce outcomes that more reflect the bureaucratic self-interest of ministries than the "national interest" (i.e., the interests of domestic economic groups and other societal actors).

We know from principal-agent theorization that political officials (ministers) as principals are in a "Yes, Minister" relationship with their civil servants in ministries (agents) (e.g., Bendor, Glazer, and Hammond 2001; Epstein and O'Halloran 1994; Strøm 2000). The executive depends on bureaucrats to translate the basic positions that they lay out into detailed and negotiable national positions, granting bureaucrats extensive opportunities to skew outcomes toward their own preferences (Peters 1995). Conversely, ministers can use centralized coordination mechanisms to control what the bureaucracy is doing (Peters 1998).

While bureaucratic politics has been the focus of much theorization, most formulations of the model are not theories per se but are instead descriptive models of decision-making processes (Bendor and Hammond 1992: 317–18; Caldwell 1977: 95; Jones 2008). Therefore, our first step is to reformulate the model of bureaucratic politics into a causal theory. The independent

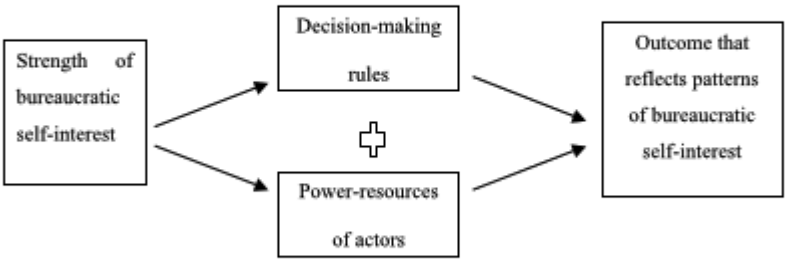

Fig. 6.2. A causal theory of bureaucratic politics

variable is the strength of the self-interest of bureaucratic actors, the intervening variables are the decision-making rules and the power resources of the actors, and the outcome should therefore be the impact of bureaucratic self-interests as mediated by the action channel used to make the decision. This is illustrated in figure 6.2.

The causal theory leads us to the following predictions that could be tested using congruence methods in a case study (Blatter and Blume 2008; George and Bennett 2005: chapter 9). First, we should expect that the strength of the dynamic of bureaucratic politics is a function of how strongly bureaucratic actors are interested in an issue. When bureaucratic actors have few self-interests in an issue (it has no real impact on the institutional power/prestige of their ministry), outcomes should reflect the national interest, and vice versa. Second, decision-making rules matter in that they determine whether the game involves coalition building around majority positions or is dominated by ensuring that the least interested actor is on board when decisions are made by unanimity (Hermann et al. 2001). Third, bureaucratic actors with procedural or hierarchical powers will be able to skew outcomes closer to their own preferred outcome. Therefore, we should expect that outcomes reflect the bureaucratic self-interest of actors as mediated by the decision-making rules and power resources of the participating actors.

For process-tracing, this causal theory then needs to be reconceptualized as a causal mechanism. However, before we do so, we briefly discuss the theoretical prior that would inform the tests of the empirical indicators in the specific case.

The Theoretical Priors

Given that the question of the importance of bureaucratic politics in the national preference formation process in EU constitutional negotiations

has not been investigated empirically by other scholars, we could utilize Moravcsik's well-established findings as the prior for the whole mechanism. Moravcsik's liberal intergovernmentalist theory suggests that the state is a relatively neutral transmission belt between the preferences of domestic economic groups and national positions: "Groups articulate preferences, governments aggregate them" (Moravcsik 1993: 483). In his major study of EU constitutional bargains, Moravcsik finds some evidence suggesting that in situations where "producer groups were balanced by strong national macroeconomic ideas and institutions," the views of governments and central banks did matter (1998: 477). However, the overall conclusion is that the form of state organization does not matter, and we can therefore treat the state as more or less a black box.

Therefore, in Bayesian terms, based on our prior knowledge, we are relatively confident that the bureaucratic politics mechanism (and its parts) are not present in this case. Updating our confidence in the presence of bureaucratic politics requires finding strong evidence that is highly unlikely to be found unless bureaucratic politics are present.

Conceptualizing and Operationalizing a Bureaucratic Politics Causal Mechanism

To develop a causal mechanism, we need to reconceptualize the causal theory into a mechanism composed of a set of parts (entities engaging in activities) to study the dynamic transmission of causal forces through the mechanism to produce the outcome (policies that more closely reflect bureaucratic self-interest than we otherwise would expect). Here, we reconceptualize the theory as a mechanism with five distinct parts: (1) preferences (bureaucratic self-interest), (2) the battle over the choice of action channel, (3) debates within group determined by the decision-making rule, (4) actors exploiting their powers (procedural or hierarchical), and (5) outcomes that reflect patterns of bureaucratic self-interest. The full mechanism is illustrated in figure 6.3. and table 6.1.

The following discussion also illustrates how the mechanism could be operationalized into a set of case-specific predictions about what evidence we should expect to find if h is valid. Here, we need to ensure that the tests for each part of the mechanism are formulated in a manner that maximizes the certainty and uniqueness of the predicted observable implications, at the same time making sure that the test is realistically feasible (i.e., data can be collected that can actually measure what we intend to measure, and so forth).

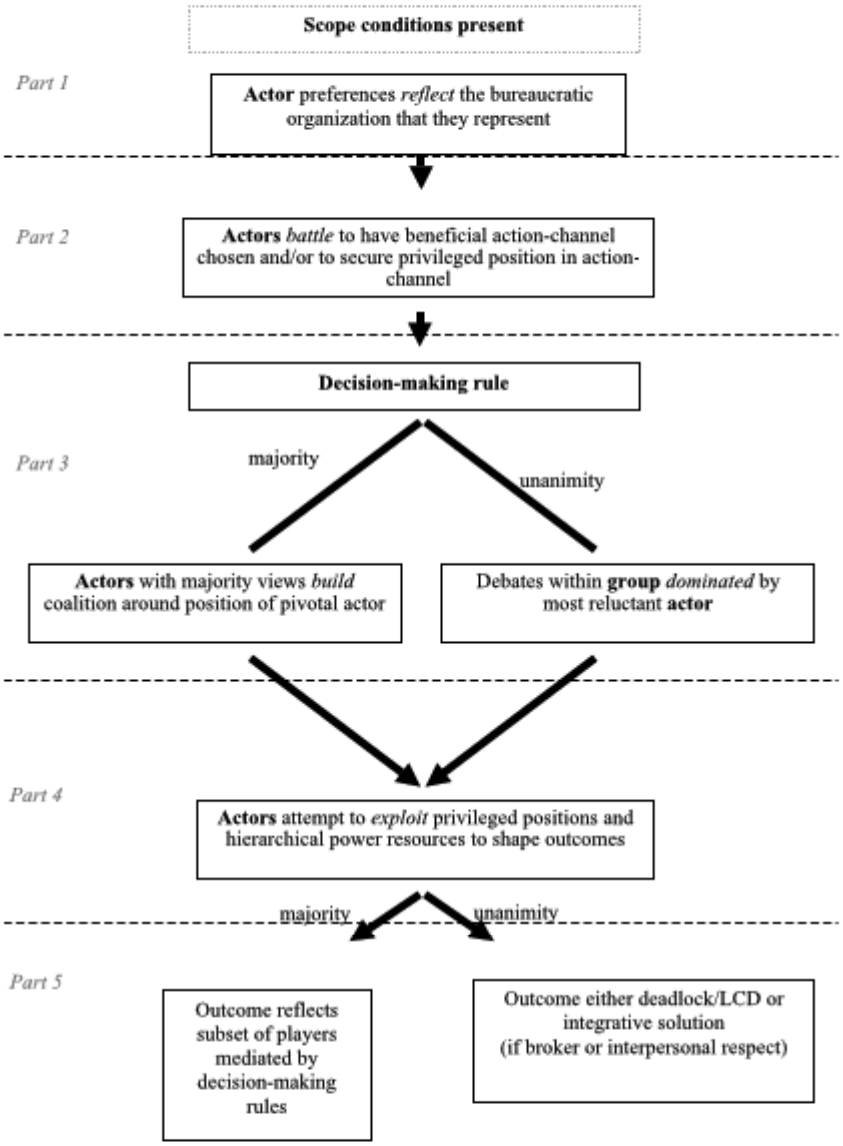

Fig. 6.3. A hypothesized bureaucratic politics mechanism for interministerial bargaining

TABLE 6.1. A Conceptualization and Operationalization of a Bureaucratic Political Causal Mechanism

	Conceptualization of Each Part	Predicted Evidence	Type of Evidence Used to Measure Prediction
1	**Actor** preferences *reflect* the bureaucratic self-interests of the ministry that they represent (competence maximizing and minimizing the disruptive effects of expected EU policy streams) (a) line ministries will attempt to *minimize* the level of disruption of EU rules upon domestic issues administered by them by minimizing or even opposing integration in their areas of competence (b) foreign ministries have interests in maintaining their coordinator position in EU affairs and therefore promote positions that reflect governmental preferences	Expect to see evidence of "bureaucratic politics talk" in the preferences of bureaucratic actors	Measured using account evidence from interviews with participants and/or internal documents produced by the ministries
2	**Actors** *battle* to have beneficial action-channel chosen and/or to secure positions with strong procedural powers in a given action-channel	Expect to see significant jockeying among actors for chairmanship of interministerial committees, etc.	Measured using sequence evidence (timing of events) and account evidence (participant accounts from interviews)
3	Debates within **group** *dominated* by either (a) (majority): Actors with majority views building coalition around position of pivotal actor (b) (unanimity): Most reluctant actor	Expect to see either (a) Minority views marginalized; coalition-building around majority view (b) Majority attempting to influence most reluctant actor to move away from SQ	Measured using account evidence from interviews with participants and/or meeting minutes
4	**Actors** attempt to *exploit* power resources (hierarchical positions or procedural powers) to shape outcomes	Expect to see actors with strong hierarchical positions and/or privileged positions in negotiation more dominant in debates and more successful in influencing outcome	Measured using account evidence from interviews with participants and/or meeting minutes

5	Outcome reflects either (a) Subset of actors determined by DM rule and privileged positions (b) Deadlock/LCD (c) Integrative solution if broker and/or interpersonal trust present in group	Expect to see that outcome reflects either (a) Subset of actors determined by DM rule and privileged position (b) Deadlock/LCD (c) Integrative solution if broker and/or interpersonal trust present in group	Measured using pattern evidence (level of congruence between outcomes and positions of ministries), account evidence from interviews with participants, and final positions produced

Part 1: Actor Preferences

The key to any theory of bureaucratic politics is the behavioral assumptions that are supposed to drive civil servants. Scholars generally believe that bureaucrats have institutional interests based on where they sit. One widely utilized behavioral assumption is that bureaucratic actors will attempt to maximize their competencies—for example, by protecting their existing turf (Dunleavy 1991; Peters 1998; Pollack 2003: 35–36). With regard to the question of the transfer of national sovereignty within EU constitutional politics, competency-maximizing national ministries can be expected to have certain interests. First, we can expect that line ministries will attempt to minimize the level of disruption of "Europe" has for domestic issues administered by those ministries by minimizing or even opposing integration in their areas of competency—in effect, protecting their turf. In some instances, however, line ministries will support more integration—primarily where further developments in the policy area are expected to result in the significant expansion of the ministry's competencies (e.g., in competition policy) or where it can be expected to result in a significant increase in the ministry's autonomy within the domestic system (Dunleavy 1991; Peters 1995). Second, we should expect that foreign ministries have interests in maintaining their coordinator position in EU affairs and therefore will promote positions that closely reflect governmental preferences to maintain governmental trust.

How, then, do we know bureaucratic politics talk when we see it in a specific case study? In other words, how can we determine whether statements from line ministries and departments are expressing parochial institutional interests rather than faithfully translating governmental preferences into positions on the concrete issues? The baseline that can be utilized is that a faithful translation of government preferences into positions should follow the overall direction of what the government has previously stated—for

example, in papers sent to Parliament prior to the start of a given set of constitutional negotiations. While most of this paper is drafted by civil servants from the foreign ministry, the key point is that it is written before the interministerial coordination process, that is, before bureaucratic politics potentially kick in. For instance, if the government stated that it is interested in transferring competencies to the EU on a specific issue but the line ministry that is responsible for the issue area opposes this transfer in study papers produced prior to a negotiation, this evidence would be an empirical indicator of bureaucratic politics talk. This could be measured using account evidence from interviews with participants both within the given ministry and in other ministries and/or internal documents produced by the ministries.

This evidence is relatively easy to gather and can provide strong evidence suggesting that the part is present/absent. This test is relatively certain in that it is necessary to observe that actors are taking positions that reflect bureaucratic self-interest so that we can be confident that the part of the bureaucratic politics mechanism exists. It is also relatively unique, as it would be difficult to otherwise explain why the line ministries most affected by EU regulation in a given issue area would also be the ones most opposed to the transfer of sovereignty on that issue.

Part 2: The Choice of Action Channel

Allison and Zelikow introduce a helpful term, *action channels,* defined as a "regularized means of taking governmental action on a specific kind of issue. . . . Action-channels structure the game by preselecting the major players, determining their usual points of entrance into the game, and distributing particular advantages and disadvantages for each game" (1999: 300–301). In this part of the mechanism, we theorize that self-interested bureaucrats will fight for the selection of an action channel that maximizes their relative powers and/or that they will attempt to secure a privileged position within a given action channel.

A test for the presence/absence of part 2 is more difficult to operationalize in a manner that permits the close observation of causal forces. We should first expect to see that actors will jockey for positions within interministerial coordination committees and that they will lobby for the selection of action channels that privilege them, other things equal. The predicted evidence of this jockeying, while difficult to gather, would, if found, update our confidence that the part of the mechanism is present. This hypothesized part

could be measured using sequence evidence (timing of events) and account evidence (participant accounts from interviews).

However, the test is not particularly certain, as there can be many reasons that evidence for this jockeying is not present. For example, there can be standing precedents for which ministry serves as chair for a specific inter-ministerial group that are sufficiently strong to overpower any bureaucratic jockeying. This suggests that this part of the mechanism is perhaps not 100 percent necessary for the mechanism to operate and implies that a reconceptualization could be considered to take account of the novelty or level of precedence for the choice of specific coordinating procedures. Further, observing jockeying does not necessarily mean that bureaucratic politics is taking place (low uniqueness); instead, it could reflect that a given ministry simply believes it is the most competent to chair a given committee based on policy expertise in the area, a phenomenon that is not evidence of bureaucratic politics. This test is therefore a smoking gun test, where finding e significantly updates our confidence in h being valid (if we can ascertain that jockeying was based on self-interest) but finding ~e does little to update our confidence in either direction.

Part 3: Debates within the Action Channel

To determine which bureaucratic actor wins the internal power struggle, Allison and Zelikow suggest that we focus on how the rules of the game structure the action channels through which bureaucratic organizations act (1999: 300–302). Unfortunately, they develop no theory for how the rules of the game determine who wins or loses. Therefore, we supplement their ideas with Hermann et al.'s theorization on the impact of decision-making rules in foreign policymaking within groups—in particular, how majority versus unanimity impacts who wins and loses (2001: 143–45).

We should expect that when majority voting is used, minority views will be marginalized unless the majority respects a minority that has an intense view, with debates focusing on a subset of options that reflect the majority (Hermann et al. 2001). If the group's members have a high degree of interpersonal trust and respect, they will attempt to bring the minority on board through integrative measures such as the use of compromises, trade-offs, and innovative solutions. Conversely, if unanimity is required, we should expect either deadlock/lowest-common-denominator solutions, as debates revolve around getting the most reluctant actor on board, or a more integrative solution in circumstances where a broker and/or interpersonal trust is present.

Operationalizing this part and the next part of the hypothesized mechanism is troublesome. The basic problem is that problems of access mean that it is usually very difficult to observe internal negotiating dynamics—it is difficult to formulate strong predictions for what evidence we should find if the parts of the mechanism exist. Instead, we are forced to rely on indirect and/or incomplete measurements. For example, while we may possess the minutes of interministerial meetings (in a best-case scenario), it is also possible that the formal institutional structures did not reflect the real workings of a negotiation. For example, it is possible that a smaller ad hoc group of influential actors met behind the scenes, fleshing out a deal in a manner that does not reflect the formal deliberations that took place in the committee and that we can observe with our data (meeting minutes). Furthermore, we risk conflating the actions undertaken by actors with influence, and we risk mistaking an actor who speaks loudly with an actor who is influential.

This said, we can operationalize part 3 by stating that we should expect that when majority decision making is used, minority views will be marginalized (except when the intense views of a minority are respected) and that we will see coalition-building dynamics taking place around a position held by a majority. Alternatively, when unanimity is used, we should expect to see discussions that are dominated by efforts to ensure that the most reluctant actor is on board. The tests formulated in this manner would be relatively unique in that these types of dynamics would be hard to explain if, for example, we theorize that foreign policy decision making is a rational and neutral process of determining the optimal solution to a given problem. However, the tests are not very certain as a consequence of the substantial empirical challenges in studying interministerial decision making; therefore, these tests are close to smoking gun tests.

It can be expected that there will be a divergence between the meetings portrayed by official minutes (which will probably not display the expected dynamics) and deliberations in the corridors between key actors (which probably will display the expected dynamics). As it is easier to collect information about what happened in official meetings than informal meetings, it can be very difficult to formulate this test with any degree of certainty unless evidence of sufficient quality is available to allow us to measure the following predictions: "We expect to see majority dynamics within the key decision-making forum (regardless of whether it is the official meetings or informal corridor deliberation)." Here we would need extensive account evidence, both from internal meeting minutes and from in-depth interviews with participants. This part could be measured using account evidence from interviews with participants and/or meeting minutes.

Part 4: Actors Attempt to Exploit Their Positions (Procedural or Hierarchical)

The impact of hierarchical and procedural positions is undertheorized in the bureaucratic politics literature (Jones 2008). Here, we tentatively theorize that hierarchy in government also determines who wins and loses, in that, for example, a minister should be able to trump the views of an undersecretary of another ministry, other things equal. Second, drawing on rational institutionalist theories, we should expect procedural powers such as chairing an interministerial committee would grant the holder certain power resources that can be exploited (Peters 1995; Tallberg 2006). For example, a coordinating ministry has responsibility for writing the final recommendation for a position, a role that can be subtly (or not so subtly) exploited to push the ministry's own interests.

Regarding part 4, we should expect to see that actors with strong hierarchical positions and/or privileged positions in negotiation more dominant in debates and more successful in influencing outcome. This part could be measured using account evidence from interviews with participants and/or meeting minutes.

However, we cannot automatically conclude that if we do not find evidence of actors attempting to exploit their power resources that they did not attempt to do so behind closed doors. In other words, any realistic operationalization for part 4 has a relatively low degree of certainty, as it is very difficult to observe this informal, behind-the-scenes use of power resources. We often must infer the process from the outcome—for example, arguing that because actor X had a disproportional impact, he/she must have successfully exploited the chair position. Here, there is the significant risk of both false negatives (actions that were not observed being mistaken for nonpresent actions) and false positives (actions being conflated with influence).

Part 5: Outcomes Reflect Bureaucratic Self-Interests of Actors

Finally, we should expect that the outcome reflects the preferences of (1) a subset of bureaucratic actors as mediated by the relative procedural and hierarchical power resources of actors, (2) deadlock/LCD under unanimity, or (3) an integrative solution if a broker and/or interpersonal trust and respect were present in the interministerial group.

Comparing the outcome with what we would expect based on the institutional self-interest of bureaucratic actors as mediated by the decision-

making rule and actor power and initial governmental positions should give us a final indication of whether bureaucratic politics was actually taking place. This test has both a high degree of certainty (in that it must be present for us to conclude that bureaucratic politics took place in the given case) and a high degree of uniqueness (as it is difficult to find any other explanation for an outcome that reflects bureaucratic self-interests of individual ministries as mediated by the action channel). The predicted evidence could be pattern evidence, such as the level of congruence between outcomes and positions of ministries, and account evidence from interviews with participants along with the final positions.

The Challenge of Conceptualizing and Operationalizing Causal Mechanisms

The basic question explored here was whether we can conceptualize and operationalize a bureaucratic politics mechanism that enables us to undertake a strong theory test in the case of national position development in EU constitutional negotiations. The tentative answer is that while we can formulate relatively strong (close to doubly decisive) tests for parts 1 and 5, we cannot formulate very strong tests for the parts in between. Instead, these are mostly smoking gun tests that attempt to assess predictions that are very difficult to measure in practice.

We conclude that we would have substantial difficulties in testing the mechanism unless we have access to both extensive primary documentation and a series of participant interviews collected either during or immediately after the negotiations (Caldwell 1977; Clifford 1990). We are therefore forced to infer that bureaucratic politics took place from evidence of parts 1 and 5 and circumstantial evidence of parts 2, 3, and 4.

In contrast, when an issue is highly salient (especially a crisis such as the Cuban Missile Crisis), extensive secondary and primary documentation is often available, and participants can recall events even many years after they took place, enabling us to access evidence that allows the analyst to observe a causal mechanism more directly. But these are not cases of bureaucratic politics but instead are governmental politics, where battles involve high-level officials (president and ministers). Here, preferences are not based on the idea that where you stand is where you sit, and negotiations have different dynamics.

The broader lesson of this example is that in many practical research situations, it is difficult to live up to this book's ideal-typical prescriptions re-

garding the conceptualization and operationalization of causal mechanisms. Most research situations are not like the one faced by Owen when studying the democratic peace, where already extensive theoretical work had already been undertaken on which he could base his conceptualization and where an encompassing empirical record was available. As the example in this chapter shows, conceptualization is challenging when theories are less well developed, and empirical testing (operationalization) can be nearly impossible when we lack full access to the empirical record. Our inferences about mechanisms are only as strong as the weakest test of a part of a mechanism.

This does not mean we should give up the endeavor when we face difficulties, merely that we should be realistic and accept that when we are using process-tracing methods, our conceptualizations and operationalized empirical tests will be less than perfect. But we should continue to strive to push our tests in the direction of the ideal.