CHAPTER 7

Turning Observations into Evidence

"There is nothing more deceptive than an obvious fact."
—A. C. Doyle 1892: 101

Empirical material needs to be evaluated before it can be admitted as evidence on which to base causal inferences. But how can we know that what we have observed is the evidence that our theory tests had predicted? How can we assess the inferential value of individual pieces of evidence? This chapter deals with the evaluation process, where raw empirical observations are assessed for their content, accuracy and probability, enabling us to use them as evidence to update our degree of confidence in the presence of the hypothesized causal mechanism. This process needs to be both transparent and open to scrutiny to ensure that it is as objective as possible.

We use the term *observation* in this book to refer to raw data prior to the evaluation of its content and accuracy (i.e., degree of measurement error). After it has been assessed for its content and potential measurement error, we use the term *evidence,* indicating that it contains a certain level of inferential value.[1]

While numerous methodological texts discuss challenges relating to the collection and analysis of data, this chapter goes a step further by making these prescriptions compatible with the Bayesian logic of inference used in process-tracing, giving us a set of evaluative tools that can be used to assess our degree of confidence in the accuracy and content of the evidence collected in process-tracing research. The prescriptions are most applicable for theory-testing and explaining-outcome variants of process-tracing as a consequence of the development of predictions about what evidence we should find in the case, capturing the need to match predicted evidence with found evidence.

The overall process of evaluating evidence is in many respects analogous to how evidence is admitted and evaluated within the U.S. legal system.[2] In a court of law, prosecutors and defenders produce different observations that can be used to make inferences about what happened. These observations include witness accounts, technical examinations, DNA tests, and so on. In a given court case, not all of the collected observations are accepted as evidence, since they can be inaccurate and/or the evidence can be so likely that it provides little inferential leverage in making an argument plausible (V. Walker 2007).

Before observations can be used in a court case, their accuracy and the probability of evidence must be evaluated. If we are dealing with a fingerprint match, we would treat a smeared print left on a surface at the scene of the crime, such as a train station, as potentially much less accurate than a print where clear ridges could be detected that would better enable comparison with a "rolled" print taken from the suspect. Regarding the probability of the evidence, a strong match between a DNA sample found on the victim and the DNA of the suspect is highly unlikely unless the two samples are the DNA of the same suspect. Combined, an accurate measure that is highly unlikely (p(e) is low) would have strong inferential weight in a court case, enabling the judge to infer with a reasonable degree of confidence that the suspect most likely was the culprit.

Taken as a whole, the judge's role is evidence evaluation—deciding which evidence is relevant and evaluating the accuracy and the probability of the evidence (V. Walker 2007: 1696). This means that a prosecutor cannot just show up in court with a gun and postulate the suspect used it to perpetrate a murder. To be admitted by the judge as evidence of a theory that a given suspect committed the crime, forensic material and/or testimony must either establish beyond a reasonable doubt or makes it highly plausible that (1) the victim was killed by a gun, (2) the gun was actually in the possession of the suspect at the time of the crime, and (3) the gun was the weapon used in the murder. In this evaluation process, the defense questions whether sources of measurement error raise doubt about the accuracy of the evidence. For example, if the testimony linking the weapon and the suspect comes from the suspect's estranged wife, should we even admit the observation as evidence when she has a revenge motive that raises serious doubts about the veracity of her testimony? Further, what is the probability of the found evidence? For example, what is the range of error of the ballistics tests used to link the weapon with the crime? Could the matching results of the ballistics tests be the product of random chance?

Similarly, in a process-tracing case study, we need to assess the content

and accuracy of our empirical observations before they can be admitted as evidence with which to update our confidence in the presence or absence of causal mechanisms; we then need to assess the probability of the evidence. The analyst needs to act as judge, prosecutor, and defense, providing arguments for why we can be confident about the content and accuracy of our observations while critically assessing our level of confidence.

Evaluating evidence involves four distinct steps that should be transparently described in one's research. Based on the predictions for what type of evidence we should expect to see if the hypothesized causal mechanism is present, we collect empirical data. Here, we need to evaluate whether we have enough data. More is not necessarily better. Instead, we need to collect strategically observations that allow us to assess our empirical tests. Further, we need to be realistic and be aware of resource limitations.

Second, we assess the content of our collected observations, using our contextual knowledge to determine what our observations tell us in relation to what evidence was predicted to occur.

Third, can we trust that the observations we have collected are in fact evidence of what we intended to measure? That is, are they accurate? Is the observation what it purports to be? What are the potential sources of error, and can we correct for them so that we can use the observation as evidence in a theory test to update the posterior probability of the hypothesized causal mechanism? This involves evaluating our confidence in the accuracy of our measure in terms of the estimated probability that the measure is accurate, depicted in Bayesian logic as p(a). We discuss the risks that nonsystematic and systematic measurement error in our collected evidence pose for our ability to update the posterior probability of a hypothesized causal mechanism.

Finally, when using the Bayesian logic of inference, we can make stronger inferences when the evidence is highly unlikely. We therefore have to evaluate the probability of the evidence by itself. This process was illustrated in chapter 6, where we saw that the more improbable the evidence (p(e) low), the stronger our ability to update our confidence in the posterior probability of the validity of a theorized mechanism when we find e. The probability of specific pieces of evidence (p(e)) therefore needs to be assessed using our knowledge of the context of the particular case.

This chapter proceeds as follows. The first section discusses some of the common challenges that all sources of empirical material share, focusing on the four stages of evaluation (collection, content evaluation, assessment of accuracy, and probability of evidence). We focus in particular on the assessment of accuracy and probability of evidence. This is followed by a presenta-

tion of the most common forms of empirical sources used in process-tracing case studies, ranging from participant interviews and archives to newspaper accounts. For each of these types of sources, we present the most common forms of measurement error and offer practical suggestions that can reduce these risks, thereby also increasing our confidence that our observations are measuring evidence that enables us to update the posterior probability in one direction or another. Different types of evidence (pattern, sequence, trace, account) can come from different types of sources. For example, account evidence can be gathered from both secondary sources (newspapers) and primary sources (participant interviews).

7.1. How Observations Are Turned into Evidence

> "The facts . . . are like fish swimming about in a vast and sometimes inaccessible ocean. . . . What the historian catches will depend partly on chance, but mainly on what part of the ocean he chooses to fish in and what tackle he chooses to use."
>
> —Carr 1961: 26

The Collection of Observations

The collection of empirical observations is not a random, ad hoc process but should instead be steered by theory, focusing on testing whether the predicted evidence for a hypothesized part of a mechanism is present in the empirical record. In process-tracing, we deliberately search for observations that allow us to infer whether the hypothesized part of a causal mechanism was present.

For example, if we are testing a hypothesized mechanism dealing with supranational entrepreneurship by the EU Commission, an empirical test of one part of the mechanism could be that we expected to find trace evidence that the Commission tabled many proposals relative to governments during a given negotiation. Relevant observations would be all of the Commission and governmental proposals during the negotiations.

It is vital to understand that we are not merely cherry-picking observations that fit a favored hypothesis when we do research. We are attempting to test whether the predicted evidence is present, meaning that we do not just go out and try to find supporting evidence; instead, we strategically seek to collect empirical material that would enable us to determine whether the predicted e or ~e is present. We therefore collect observations that enable us to put our hypothesized causal mechanism to a critical test (usually in

the form of hoop tests, but in certain circumstances double-decisive tests). The researcher has to evaluate, at each stage of the process, whether the data collected is enough, whether the residual uncertainty is acceptable, and whether additional data should be collected to find sufficient evidence for the existence of e or ~e.

The collection of observations (either primary or secondary) always incurs the potential risk of bias as a consequence of the nonrandom selection strategy used. This might result in skewed inferences that are based on a sample that is not representative of the population of potential evidence (Collier 1995: 462). According to Collier and Mahoney, selection bias occurs when "some sort of selection process in either the design of the study or the real-world phenomena under investigation result in inferences that suffer from systematic error" (1996: 58–59).

This danger is particularly acute in process-tracing research. Sources often select the researcher—for example, where availability concerns determine which sources we use (Thies 2002: 356). There can be reasons why a particular record has survived through time, whereas others might have been deliberately destroyed. Further, our access to relevant sources might be restricted, limiting our ability to get the material we need to collect to test our hypotheses in an unbiased fashion.

Selection bias might also occur even when secondary rather than primary sources are used. Lustick (1996) has suggested that a researcher's choice of a particular historical monograph can be seen as holding the potential for selection bias. A historian may misinterpret the evidence, drawing incorrect inferences about past events, or may leave out important facts about the event, making the account unfair or unbalanced. Another concern is the selection effects of the social scientist when she chooses to focus on a particular historian's work and consciously—or unconsciously—excludes the work of others (Larson 2001; Lebow 2001). In the worst-case scenario, the two problems are combined, and the social scientist with a particular theoretical and conceptual predisposition purposefully selects certain historians who share this bias and whose work is already tainted, producing faulty confirmation of the social scientist's theory (Lustick 1996: 606). In most cases, we should not rely on a single historical work but rather should try to incorporate different historians who belong to different historiographic schools in our analysis (Lustick 1996).

Testing whether a causal mechanism is present involves assessing whether our theory-based predictions of what evidence we should see in the empirical record are matched in reality. The predicted evidence can take the form of pattern, sequence, trace, or account evidence. Relevant evidence depends on what empirical manifestations were predicted by the tests. Collecting differ-

ent types of evidence raises different types of challenges. Pattern evidence in the form of a number of proposals in a negotiation can be time-consuming to collect, and important holes can exist when documents are unavailable for one reason or another. Conversely, collecting accurate elite interviews that provide account evidence about what happened in a given negotiation raises numerous challenges regarding access to high-level decision makers and regarding the independence of their accounts. In collecting data for process-tracing, the most appropriate analogy is that of a detective, with the collection of relevant evidence involving the expenditure of much shoe leather (Freedman 1991).

Finally, the collection of empirical material should be seen as a cumulative process. Even with the most strenuous efforts, the evidence we gather can always be improved in subsequent studies. Drawing on an analogy with astronomy, until recently, the theory that other solar systems had earth-like planets was merely a hypothetical conjecture. Recent advances in telescope technology have enabled the collection of evidence showing the existence of earth-like exoplanets. Future technological developments will allow for more accurate measurements to be taken that will increase our confidence in the theory. Similarly, in process-tracing research, evidence gathered will always have a preliminary form, and the results of our research (the posterior probability) can always be updated by new evidence. Important archival evidence can be declassified, memoirs are published, and scholars who critically analyze existing evidence can develop more accurate interpretations of e that either increase or decrease our confidence in the presence of hypothesized causal mechanisms. Therefore, scholars using process-tracing methods need to be as aware as historians that any result can be updated when new sources are found that better measure e. In the words of Elman and Elman, "Historians know that there are likely to be other documents, indeed whole collections of papers that they have not seen. Accordingly they are inclined to view their results as the uncertain product of an incomplete evidentiary record" (2001: 29).

Assessing the Content of Observations

Once we have collected observations, we need to evaluate critically what those observations tell us in light of our background knowledge about the context. This is the first phase of evaluating whether our observations are evidence ($o + k \rightarrow e$). The next phase involves assessing accuracy.

What does the observation tell us? What is the source of the observation and the context in which it was produced? Answering these questions

requires considerable background knowledge. For example, how does the given political system work? Is there anything amiss in the events that have been uncovered? What is normally included in the type of source (e.g., minutes of a cabinet meeting)? It is important to interpret documents within their historic, situational, and communication contexts. We need to understand the purpose of the document and the events leading up to it to interpret its meaning correctly (Larson 2001: 343). According to George and Bennett, the analyst should always consider who is speaking to whom, for what purpose, and under what circumstances (2005: 99–100). The inferential weight of what is contained in a document often cannot be reliably determined without addressing these questions. Similarly, when conducting elite interviews, it is vital to consider whom is being interviewed and his or her motivations for doing so.

This means that each observation should be evaluated relative to what is known about the actors, their intentions, their interactions, and the situation in which they found themselves (Thies 2002: 357). When analyzing documents, one must ask what purpose the document was intended to serve and what agenda the author might have. How did the document fit into the political system, and what is the relation to the stream of other communications and activities within the policymaking process? It is also important to note the circumstances surrounding the document—why has it been declassified or released and for what purpose—and what has not been released?

Evidence might also be what is not mentioned in the text or in an archive. This type of evidence can be called *e silentio* evidence—based on silence or the absence of an expected statement or message in a text or in an archive. If we expect an author to have a strong interest in presenting an event in a certain light or we would expect the author to take credit for a given decision, and such is not the case, this omission might have inferential weight in our analysis. When a certain event is not mentioned in the text, one possible explanation might be that the event did not take place. Conversely, if we are very confident based on other sources that an event took place, its omission in a source could be a highly improbable piece of evidence with strong inferential weight. As the "dog that did not bark" example showed us, silence or the absence of something can in certain circumstances be the strongest evidence.

Evaluating the Accuracy of Observations

The next stage of the evaluation process is to assess its accuracy. In the measurement of social phenomena, we can never aspire to 100 percent accurate

measurements, but through a critical evaluation of the size of measurement error, we can increase our degree of confidence that our collected observations are what we think they are.

Is the observation what it purports to be? Are we measuring what we intended to measure? Inaccurate measures can be the product of either nonsystematic or systematic errors in the measuring instrument we use to collect observations. Nonsystematic (or random) error is commonly termed *reliability,* whereas systematic error is defined here as the *bias* in our measuring instrument. For example, a telescope that has random imperfections in its mirrors will produce a blurrier image. The result would be haphazard pictures that prevent us from making accurate observations regarding the exact location of stellar objects. However, the error is random. In contrast, bias is the product of a measuring instrument that produces systematic errors. If there is an error in a mirror that results in the image being skewed systematically two millimeters to the left, this is a much more serious form of error, as it distorts the picture in a systematic pattern, resulting in invalid inferences.

These two forms of measurement error and the risks that they pose for causal inferences can be expressed in Bayesian terms. Unreliable measurements decrease our confidence that we have actually measured e, which then reduces the ability of our empirical tests to update our confidence in the presence/absence of a part of a mechanism (the posterior probability). Accuracy can be expressed in terms of probability (p(a)), where an unreliable measure has a low probability of being accurate, and vice versa. Entered into Bayes's theorem, an unreliable measure reduces the ability of evidence to update our confidence in whether or not a hypothesis is true.

Howson and Urbach provide a Bayesian reasoning for this statement (2006: 111). The probability of accuracy (p(a)) enters the calculation of the posterior probability of a hypothesis through the likelihood function. The nominator of the likelihood ratio is p(e|-h), which logically equals p(e|-h & a)p(a) + p(e|-h & -a)p(-a). The latter means that the probability of e being found if the hypothesis is not true when the measure is accurate is multiplied by the probability of the accuracy of the instrument. This sum is then added to the product of the probability of e being found if h is not true and the instrument is not accurate multiplied by the probability that the instrument is not accurate. In the denominator, p(e|h) = p(e|h & -a) × p(-a) means that the probability of e being found if the hypothesis is true equals the probability of e being found if h is true when the measure is inaccurate times the probability of the measure being inaccurate. Taken together, these calculations show that a very unreliable measure increases the size of the denominator and decreases the nominator in the likelihood ratio. If we find e but p(a) is low, finding e does little to update our confidence in the verac-

ity of the hypothesis, meaning that the posterior probability is not updated substantially.

The best solution to the problem of unreliable measures is to collect multiple independent observations. This approach is commonly referred to as *triangulation* and can refer to collecting observations either from different sources of the same type (e.g., interviewing different participants) or collecting observations across different types of sources (e.g., archives and interviews) or different types of evidence, if available (i.e., pattern, sequence, account, trace). However, triangulation does not help unless we can substantiate that the sources are independent of each other. Doing three interviews and postulating that sources have been triangulated is not enough—the researcher needs to substantiate the fact that the interviews are independent of each other.

The importance of the independence of sources can be understood using Bayesian logic. If pieces of evidence are truly independent, it is highly unlikely that the measures will result in the same evidence (e) unless e is actually a true measure (Howson and Urbach 2006: 125). If we are conducting interviews to measure what happened in a political negotiation and three participants in the negotiations offer similar accounts, and if we can verify that the accounts are independent of each other, we would increase our confidence in the accuracy of the evidence of the account of the negotiations. It would be highly unlikely to find similar accounts unless the observed e is a true measure. However, finding similar accounts could also mean that the participants met afterward to agree on a common account of the events. If they had met, finding that the accounts were similar would do little to update our confidence in the accuracy of the evidence. And finding that the accounts are too similar should actually decrease our assessment of the accuracy of the measure quite dramatically, as it is highly unlikely that we would find 100 percent correspondence unless the measure is not accurate.

Biased observations mean that the error has a systematic pattern. The risk of this problem is particularly acute in process-tracing research, where the observations we collect are not a random sample. We need to be particularly concerned with systematic error when the bias is systematically related to the patterns predicted by (e|h) or (e|~h). This worst-case scenario can occur either inadvertently (when the measuring instrument produces evidence that systematically confirms or disconfirms h) or when a researcher deliberately chooses observations that either confirm or disconfirm the pattern expected by (e|h) or (e|~h). In this scenario, the evidence collected is either too good or too bad to be true (Howson and Urbach 2006: 116–18). If the found e fits the hypothesis perfectly, we should be very suspicious, as this result is im-

probable in the real world. Howson and Urbach use Bayes's theorem to show why collected evidence that perfectly fits with either (e|h) or (e|~h) does not result in the updating of the posterior probability. Rigging the results would mean that we would expect to find e regardless of whether or not h was true, meaning that the probability of finding the evidence when using a highly inaccurate measure (~a) would be high regardless of whether the hypothesis was true or not (p(e|h & ~a) = p(e|~h & ~a)). When this is the case, it can be shown that the posterior probability (p(h|e)) equals the prior (p(h)), meaning that when e is rigged, no updating of the posterior takes place.[3]

In contrast to the problem of reliability, there are no quick fixes to correct for systematic bias in our measurements. Therefore, being suspicious is a rule of thumb in evaluating observations. Observations that appear to be too good to be true are probably false. Second, corrections can be made by critically evaluating the size and direction of bias. This can be accomplished by a critical assessment of the source of each observation and by comparing the observation with other independent observations in a triangulation process to assess the size and direction of bias. We should focus in particular on assessing the degree of systematic error that either favors or disfavors the hypothesis.

Evaluating the Inferential Weight of Evidence: How Probable Is the Evidence?

Finally, once we are reasonably confident that the evidence we collected is accurate, we need to evaluate the probability that particular pieces of evidence exist, tying our evaluation into the Bayesian theorems presented in chapters 5 and 6. The term p(e) in the first form of Bayes's theorem states that the inferential weight of evidence is contingent on its probability in the context of the specific case. This refers to evidence independent of its relation to a hypothesis, whereas when the relationship between theory and evidence is considered in the likelihood ratio, this relates to considerations of test strength that were discussed in chapter 6.

We evaluate p(e) based on our contextual knowledge of particular cases. The contextual sensitivity of evidence is one of the strongest comparative advantages of in-depth, qualitative case study methods, where expert substantive knowledge of individual cases is used to evaluate what constitutes evidence and whether it is highly improbable or probable to find specific pieces of evidence in a particular case.

The probability of evidence relates directly to the question of how many

pieces of evidence are necessary to update our confidence in the validity of a hypothesis. If p(e) is very low, even one piece of evidence can be enough to update significantly our confidence in the validity of a hypothesis. For example, using an example from the *Silver Blaze* story, another part of the hypothesized mechanism explaining the crime was that Straker planned to injure the prize racehorse by cutting its tendon so that it could not run in the race. This part of the mechanism relates to how the crime was planned.

Holmes tests this part by developing a prediction that Straker would have had to practice the delicate tendon-nicking technique before he attempted it on the racehorse. The evidence that Holmes collects to test this prediction is that three of the "few sheep in the paddock" had recently gone lame. To evaluate the inferential weight of this evidence in relation to updating his confidence in the part of the mechanism, Holmes assesses the probability that three out of a handful of otherwise healthy sheep went lame in a short period of time. He concludes that the probability of finding this evidence was highly unlikely, enabling him to use the found evidence to make a strong inference about the part of the mechanism. In contrast, if sheep were known to commonly have hip problems like those that plague certain large dogs, the probability of finding the evidence (lame sheep) would have been quite high, meaning that little updating would have been possible based on the found evidence.

The key point here is one single piece of evidence, if sufficiently improbable, can significantly increase our confidence in the validity of the hypothesis. Schultz explicitly uses this type of reasoning when he presents evidence in a case study of the 1898 Fashoda crisis between France and the United Kingdom.

> As predicted . . . signals from the government were sent in a way that entailed high and visible audience costs. They were made publicly, in full view of the British electorate, and they were designed to arouse public opinion so that it would be difficult for the government to later back down. The most prominent example of such signaling came on October 10, when Salisbury took the *unusual step* of publishing a blue book on the crisis, a collection of key dispatches between the two countries. Until this point, the negotiations had taken place in private. With the publication of the blue book, the positions taken and arguments made by both sides were out in the open. The British public could see for its own eyes the uncompromising position of the government, as well as the audacity of French claims. Salisbury's action was *not only unusual but a breach of prevailing diplomatic norms:*

> "as a matter of courtesy, records of diplomatic negotiations are not generally given to the public until the negotiations with which they are concerned are ended." . . . At the same time, the move had tremendous signaling value. By publicizing Britain's stance in such an *unusual manner,* Salisbury effectively painted himself into a corner: retreat from this position would entail substantial political costs." (2001: 186–87; emphasis added)

The word *unusual* appears numerous times to hammer home the point that one piece of improbable evidence (here, making public the blue book) is enough to substantiate the claim that the U.K. government was creating high and visible audience costs for itself. Evaluation of p(e) is case-specific. What is unusual in one case might be common evidence in another. Schultz substantiates the unusual nature of publicly airing the blue book by mentioning that publishing key dispatches is something that not only deviates from normal U.K. government behavior at the time but also is a "breach of prevailing diplomatic norms." By itself, the observation is just an observation; combined with contextual knowledge, it is evidence that is highly improbable and therefore has a stronger inferential weight than more probable pieces of evidence.

All evidence is not created equal. The Bayesian logic of inference and the role that the probability of evidence plays in it implies that we should attempt to design bold predictions for finding relatively improbable evidence. However, this raises two dilemmas. First, our empirical tests must be practical. Surprising or improbable evidence is difficult to access, and if we have access, very difficult to find. Second, we need to ensure that our selected evidence is representative of general trends in the evidence. A single piece of evidence like the blue book could be an anomalous observation. Therefore, when using highly improbable evidence, it is important to establish that the single piece of evidence reflects broader trends.

Consequently, in most empirical tests, we make predictions that rely on multiple pieces and sources of evidence. However, it can be argued that utilizing multiple different pieces and sources of evidence that are independent of each other actually decreases p(e): How probable is it that we find the same confirming evidence when we utilize different sources of evidence? In a study of the Cuban Missile Crisis, finding the same evidence in U.S. and Soviet archives would be highly improbable unless our observations are measuring e, whereas finding that U.S. participant accounts in memoirs and interviews matches U.S. archival records is much more probable.

Depending on the type of evidence, very different forms of evaluating its probability can be appropriate. If we are assessing pattern evidence, classic statistical probabilities can be relevant to assess the probability of finding e given what we know about the normal distribution of populations and the like. However, even here, we are using reasoning that is more analogous to detective work. Freedman offers a good example of the evaluation of the probability of pattern evidence based on Snow's work on the causes of cholera: Snow "made a more complete tally of cholera deaths in the area. His 'spot map' displays the location of cholera fatalities during the epidemic, and the clustering is apparent from the map" (2011: 226–27). Crucial here is the evaluation of the probability of the pattern (p(e))—for example, by using statistical tests to determine the expected probability of a correspondence between distance to infested water supplies and the incidence of cholera. Conversely, if we are evaluating sequence evidence, how unusual would a given pattern of meetings be given what we know about the context?

7.2. Sources of Evidence in Process-Tracing

The selection of sources in process-tracing research is not driven by random sampling; instead, we select sources based on the type of evidence that is best suited to enable us to engage in a critical theory test. In other words, source selection is theory-driven.

Different sources of evidence are commonly used in process-tracing analysis, and each has its own benefits and pitfalls. We first introduce one way in which primary and secondary sources can be delineated, using the widely accepted Anglo-Saxon terminology (e.g., Trachtenberg 2006). We then discuss the shortcomings associated with particular sources and tools that can be used to improve the accuracy of measurements based on these sources.

The Distinction between Primary and Secondary Sources

We define primary sources as eyewitness accounts of a given process—for example, documents produced by participants at the time an event occurred. Secondary sources, in contrast, are produced based on primary sources. Thus, the work of a historian studying primary sources (e.g., the documentary record of a negotiation) is a secondary source.

It can often be difficult to determine whether a source is primary or sec-

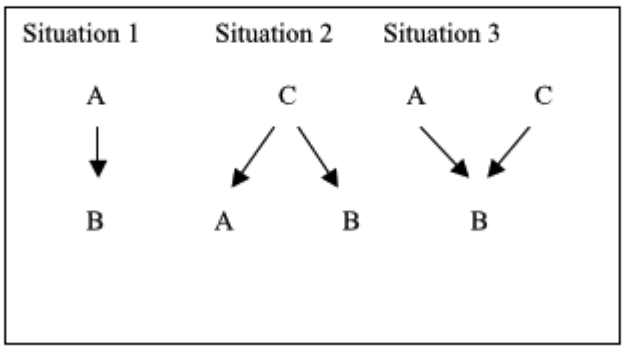

Fig. 7.1. Dependence between sources. (Based on Erslev 1963.)

ondary. An interviewee can be a primary source; however, if the respondent did not actually participate but has the information from other participants, the interview would be a secondary source.

We can use tools such as the dating of documents and text analysis to determine which sources are primary and secondary (Milligan 1979). For example, if source A has the same linguistic phrasing as source B and we can show that the production of source A preceded source B, this suggests that source B should be treated as secondary material that builds on source A, as it is highly unlikely that they would have used the same phrasing unless source B draws on source A.

Logically, there are three ways that a source A can be related to source B (figure 7.1.). First, source B has the information from source A. Here, source A is primary and B is secondary. We should therefore, other things equal, attribute higher inferential weight to A rather than B. Second, if sources A and B are reports drawn from a common source C and C is unknown for us, then A and B are both primary to us. However, they are not independent of each other given the common source. Therefore, when we possess A, adding B would not provide any new additional evidence. If C is known, then A and B should be considered secondary and C primary. We should therefore use C instead of A and B. Third, if parts of B draw on A, but B also reports something from C and C is unknown for us, then A is primary, the parts of B that draw on A are secondary, while those parts that rely on the unknown C are primary to us. However, if both C and A are known, then they are primary sources, while B is secondary (Erslev 1963: 44–45).[4]

Interviews

One of the most commonly used sources of evidence in process-tracing research is interviews, be they elite interviews, where the respondents provide information about events or their motivations therein, or interviews with persons who are treated as representatives of a particular worldview. This discussion focuses on elite interviewing, since the goal of many process-tracing studies is to gain information about political events by interviewing the key actors in the process (Tansey 2007). Interview observations are primarily used to supply account evidence, where we are interested in participants' recollections of different aspects of a process, and sequence evidence, where we want to gather information about the sequence of events that took place in a process.

Before the content of individual interviews is assessed, one needs to consider whether the selection of sources is biased in one direction. If we are analyzing a political negotiation, have we spoken to both the winners and losers? To quote a proverb, "Success has many fathers; failure is an orphan."

The account provided by an interview must be assessed. What does the respondent say? That is, what is the observation? One of the advantages of elite interviews is that they provide the opportunity to interview the persons who actually participated in the case under investigation. Participant accounts potentially offer a more direct measure of a causal mechanism, depending on how the theoretical test has been operationalized. Further, interviewing allows the researcher to move beyond the written accounts and gather information about the underlying context of events.

Once we have determined the content of the observation, we need to assess the degree of accuracy of evidence in terms of reliability and bias. The first question to ask is, What was the interviewee's role in the process? Should the observation be treated as primary or secondary material? A primary source was directly involved in an event, while the secondary source was not present but instead draws on information from other sources (e.g., other participants or the minutes of meetings).

Normally, we should expect that the accuracy of a primary source is higher than that of a secondary source, but this might not always be the case. A participant who was either unable to comprehend what was happening or who perceived events incorrectly is less accurate than a secondary account by an expert observer who had full access to the documentary record and many different participant accounts. However, if the observer is biased toward a particular theory about why events happened, the secondary account

would also contain significant measurement errors that would decrease the accuracy of the observations.

A particular challenge is raised by the fact that respondents sometimes overstate their centrality in a political process. Could the respondent have known what took place behind closed doors? Kramer gives the example of former Soviet ambassador to the United States Anatoly Dobrynin, whom some historians have used as an inside source for observations about the Soviet decision-making process leading to the Cuban Missile Crisis. However, despite Dobrynin's claims, Kramer argues that Dobrynin was not privy to the highest-level deliberations (1990: 215). Most damningly, Dobrynin was not informed about the deployment of missiles until after the United States had discovered them (215). Therefore, we would assess his account of the high-level deliberations as being less accurate, other things equal.

When we are dealing with secondary sources, the observations might be unreliable if the interviewee has relied on hearsay. In this type of interview, the researcher needs to ask the respondent about the sources for his/her claims about what happened in a political process. Does the respondent build his/her claims on the minutes of the meetings or from detailed discussions with participants immediately after the negotiations took place?

Another pitfall is the length of time between an event and when the interview is conducted. Participants interviewed immediately after a negotiation can be expected to have a reasonable recollection of the details about what took place, but lapses of memory over time will result in observations that are less reliable. More insidious is the risk that participants will, by reading other accounts and talking with other participants, change their interpretation of what happened to match other accounts, resulting in potential bias in our observations if we interview these participants long after the fact.

To assess potential bias in the observations, one needs to ask whether the respondent has a potential motive for presenting a skewed account of events? Indeed, one needs to ponder why the respondent has chosen to be interviewed. If one is interviewing only the winners in a political negotiation, this should raise warning flags about the potential bias of the material provided.

As a consequence of the imperfections of human memory, interviews will never be a perfectly reliable measuring instrument. Reliability can, however, be improved through the careful use of triangulation both across different persons and between different kinds of sources (interviews, archival observations, and so forth).

However, for triangulation to work, we need to establish that the sources are independent of each other. If we are triangulating across interviews, we

need to make sure that we have conducted interviews with participants from different sides to establish the independence of sources. If we are investigating a case involving a negotiation of a bill in the U.S. Congress, we would say that triangulation across independent sources has occurred if we interviewed members of both parties in the House and Senate, along with their aides and lobbyists who were involved, together with White House representatives who took part in the negotiations. If we find the same account evidence in multiple independent sources, it would be highly unlikely to find this unless e is an accurate measure of e.

Triangulation across different kinds of sources can also be used to check the accuracy of interview observations (Tansey 2007), again contingent on the independence of the different sources. If we interview a person who took part in a meeting and then find the same observations in the minutes of the meeting, this could increase our confidence in the accuracy of the observations. However, if the person we interviewed also wrote the meeting minutes, finding the same observation in two different kinds of sources would do nothing to increase our confidence in the accuracy of our observations.

Archival Material

Process-tracing scholars often aspire to produce case studies that build on what social scientists have often termed "hard" primary evidence—for example, official internal documents produced by public authorities that describe what took place behind closed doors (e.g., Moravcsik 1998: 80–85). While even internal documents can include measurement error, when they are generated as a nonpublic record of what took place, we can be reasonably confident that they are accurate: As Trachtenberg notes, "What would be the point of keeping records if those records were not even meant to be accurate?" (2006: 147). Archival material can provide all four types of evidence. Pattern evidence could, for example, entail counting the number and length of documents. Sequence evidence could be in the form of documents, describing what meetings took place over a period of time (e.g., agendas for meetings). Trace evidence could take the form of meeting minutes as proof that a meeting actually took place (a trace). Finally, meeting minutes could also be used as account evidence for what took place there.

The first step in evaluating the accuracy of archival material is to examine the authenticity of a document. If something seems amiss in the authorship, time period, style, genre, or origin of a document, we have to uncover its past to evaluate whether the source can tell us anything. Relevant ques-

tions are: (1) Has the document been produced at the time and place when an event occurred, or has it been produced later and/or away from where the events took place? (2) Is the document what it claims to be? (3) Where and under what circumstances was it produced? (4) Why was the document created? and (5) What would a document like this be expected to tell us? Naturally, if the document is not authentic, it has no inferential value for us.

The rest of this section deals primarily with challenges related to dealing with official archives, such as foreign ministerial archives. However, for many of the research questions in which social scientists are interested, this type of archive can be either uninformative or irrelevant. The personal archives of participants can also be relevant, but daunting challenges of access can arise. For many research questions, there are no relevant archives, meaning that the historical record must be traced using other types of sources.

The historical record is usually quite ambiguous (Wohlforth 1997), and we do not suggest that ambitious scholars enter the archives before they have operationalized strong tests of the presence/absence of each part of the hypothesized causal mechanism. Finding answers in archival records is often akin to finding a needle in a haystack. Without a clear set of theory tests to guide one's search in the archives, it is more akin to finding some small object in the haystack without knowing what that object is. The utility of using a theory-driven search for data is illustrated by an exchange from the Sherlock Holmes story *Silver Blaze*:

> Holmes took the bag, and descending into the hollow he pushed the matting into a more central position. Then stretching himself upon his face and leaning his chin upon his hands he made a careful study of the trampled mud in front of him.
>
> "Halloa!" said he, suddenly, "what's this?"
>
> It was a wax vesta, half burned, which was so coated with mud that it looked at first like a little chip of wood.
>
> "I cannot think how I came to overlook it," said the Inspector with an expression of annoyance.
>
> "It was invisible, buried in the mud. I only saw it because I was looking for it." (A. C. Doyle 1975: 17)

Before we can admit observations gathered in archival material, we need to assess what the observation is in light of our background knowledge (e.g., the context in which it was produced) and the potential sources of measurement error that can exist in the document.

Basic questions include the document's source. Can we assume it is au-

thentic? Why was the document created, and what were the circumstances and motives surrounding its creation (Thies 2002: 357, 359)? Trachtenberg (2006: 141) gives examples of questions one could ask to put documents in context in a particular international conflict: What did each country want? What policies were they pursuing? In what kind of thinking was the policy rooted? What did each side actually do? How does the document fit into this story? Here, we need to assess the purpose of the document and the empirical process leading to its creation to enable an accurate interpretation of what evidence the observation provides (Larson 2001: 343). That is, we need to evaluate $o + k \rightarrow e$.

What does the source tell us? Can it inform us about the part of the causal mechanism we are examining? Is the evidence a trace of a hypothesized part of a causal mechanism, where the mere existence of given piece of evidence substantiates that a hypothesized part of the mechanism existed? For example, the existence of an internal study paper could substantiate the claim that the government put at least some effort into a negotiation. Or is the source an account of events that allegedly took place, like the minutes of a meeting? In addition, was the source produced when the event took place or much later? Did the author personally take part in the event, or was the document produced based on the accounts of others? For example, diplomatic cables often report secondary accounts of meetings or contain a leader's interpretation of a situation as recounted by one of his aides.

In assessing archival sources, it is useful to consider what this kind of document usually contains. This can be termed the genre of a document. What is usually included in the minutes of a National Security Council meeting or a CIA study paper?

After we have assessed the observations in light of what we otherwise know about the genre of document and the context in which it was produced $(o + k)$, we need to evaluate whether the observation has any plausible measurement error. Does the source give a reliable account of what it alleges to measure (Trachtenberg 2006: 146)? If the source alleges that it records what took place in a meeting, does other evidence also show that the meeting was held (146)? If the meeting was between governments, do we find accounts in the archives of the other governments (146)? If the document was produced much later than the meeting, we should, other things equal, expect it to be less reliable.

Do the observations contain any form of systematic error? We know that high-level actors attempt to skew the historical record to favor their own accounts (Wohlforth 1997). One way in which this can occur is that

documents are selectively declassified, with the sources available all slanted toward the account of events favored by authorities. We therefore have to ask ourselves why this particular document has been declassified while others have not. Does this selective declassification have potential bias?

Further, what is the likelihood that the observations themselves have been manipulated? In other words, do the minutes of meetings reflect what took place or the account favored by a decision maker? The risk of this form of bias is particularly problematic in centralized dictatorial systems such as the Soviet Union, where official records only reflect what leaders want them to say (English 1997). In the Soviet archives, there are strong indications that the documentary record reflects the views of leaders, especially during the Stalin years.

But the problem also exists in democratic systems. The BBC TV series *Yes, Prime Minister* illustrates the problem in a (hopefully!) exaggerated example of a fictional exchange between two civil servants, Bernard (B) and Sir Humphrey (SH), regarding the production of minutes of cabinet meetings:

> SH: What I remember is irrelevant. If the minutes don't say that he [the Prime Minister (PM)] did, then he didn't.
> B: So you want me to falsify the minutes?
> SH: I want nothing of the sort! . . .
> B: So what do you suggest, Sir Humphrey?
> SH: The minutes do not record everything that was said at a meeting, do they?
> B: Well, no of course not.
> SH: And people change their minds during a meeting, don't they?
> B: Yes . . .
> SH: So the actual meeting is actually a mass of ingredients from which to choose from. . . . You choose from a jumble of ill-digested ideas a version which represents the PM's views as he would, on reflection, have liked them to emerge.
> B: But if it's not a true record . . .
> SH: The purpose of minutes is not to record events, it is to protect people. You do not take notes if the PM says something he did not mean to say, particularly if it contradicts something he has said publicly. You try to improve on what has been said, put it in a better order. You are tactful.
> B: But how do I justify that?
> SH: You are his servant.
>
> (BBC 1987)

Another challenge with archival documents relates to intelligence estimates and reports from embassies. Here, we need to think of the relationship between the producer of the document and the consumer. A classic example occurred in the early 1960s during the Vietnam War, where assessments of the situation produced by junior CIA officials in the field were consistently very pessimistic (Ford 1998). However, more senior CIA officials were cognizant of what the consumers of intelligence in Washington, D.C., wanted to hear. Therefore, they pressured junior officials to "get with the team" and edit documents in a more optimistic direction. If we are testing a part of a causal mechanism dealing with whether high-level decision makers knew the real situation on the ground in Vietnam, our results would be very different depending on whether we gain access to the first, unedited pessimistic version or the redacted final optimistic version sent to Washington, D.C.

As with interviews, triangulation can be used to assess and potentially to correct for measurement error, contingent on the different sources being independent of each other. If we find the same account of an event in the archives of two different governments, it is highly unlikely that we would find these two observations unless they actually are measuring what took place.

Memoirs, Public Speeches, and Other Forms of Primary Sources

We now turn to a variety of primary sources that are generally treated as "soft" primary sources—the memoirs and diaries of participants, private letters, and public statements and speeches. Common to these primary sources is that they are usually intended to be made public. Only the most naive person would expect that private letters that detail important events will remain private after one's death. Therefore, even private letters can be expected to be written with the understanding that they might be made public at some point. We can, however, use these sources to provide account evidence—taken with more than a few grains of salt. In contrast, we can usually trust these softer sources for sequence evidence, although in particularly sensitive negotiations, there also can be strong motives for participants to distort even the timetable of events.

Memoirs are intended to be made public, and therefore are only an accurate picture of what an actor wants others to believe happened. Private diaries and letters of participants in historically important events can also be expected to be biased as a consequence of the high probability that they will someday be made public. Like memoirs, they therefore are an accurate measure only of the participant's take on events for public consumption.

When assessing these sources of observations, we need to ask the same questions that we ask of interview observations: How close was the person to the events? Did the person participate, or is the information from hearsay? Does the source have a motive to present an accurate account of the events? For example, politicians might have a tendency to overstate their importance in a series of events, whereas civil servants often downplay their roles to maintain the appearance of neutrality.

We should be particularly cautious in claiming that an observation is accurate when the information contained in it aligns with the interests/motivations of the source. Conversely, if the author gives an account that is against his or her interests, we are not likely to find this observation unless it is accurate, meaning that we can attach more confidence to its accuracy.

Public statements and speeches can be used as evidence in specific circumstances. Speeches often are used to justify policy choices and therefore cannot be used to measure the real motivations behind a decision. For example, public justifications for war in Iraq included speeches by Bush about weapons of mass destruction that were not necessarily the real motivation behind the invasion. However, if we find the same justifications in private settings where policymakers can speak more freely, it is unlikely that we would find the same statements in both the public and private record unless they accurately reflect policymakers' motivations (Khong 1992: 60).

Historical Scholarship

When we are doing process-tracing research, we are usually very dependent on secondary historical scholarship. In the words of Skocpol, "Redoing primary research for every investigation would be disastrous; it would rule out most comparative-historical research. If a topic is too big for purely primary research *and* if excellent studies by specialists are already available in some profusion—secondary sources are appropriate as the basic source of evidence for a given study" (quoted in Lustick 1996: 606). Historical work can be used as sequence and account evidence.

When utilizing historical reports, we must first remember that historical work is not "the facts." We need to assess carefully the reliability and potential bias of each observation that we use from historical scholarship.

Historians are people, too, meaning that they can make mistakes. Historians may misinterpret primary sources, resulting in incorrect descriptive inferences about what happened. This does not mean that we cannot use historical accounts but merely that we must be aware of the fact that any

given work is potentially unreliable. To reduce this risk, we should triangulate observations from multiple historical sources to ensure that we reduce the risk of random measurement error.

More problematic is the fact that the work of historians reflects their implicit (or sometimes explicit) theories. The "historical record," therefore, is not a neutral source of information but instead reflects these theories (Lustick 1996). There is a substantial risk of bias when a social scientist with a particular theoretical and conceptual predisposition purposefully selects work produced by historians who share this bias, resulting in the mistaken confirmation of the social scientist's preferred theory (606).

For example, in scholarship on the start of the Cold War, the first generation of work in the 1950s tended to favor the theory where the Soviet Union and either the innate expansionistic tendencies inherent in communist doctrine (the "neurotic bear") or Soviet behavior as a traditional great power were the cause of the Cold War (e.g., White 2000). This school was followed by the revisionist accounts in the 1960s that pinned blame on the United States, contending that U.S. expansion as a consequence of capitalist economic interests triggered the Cold War (e.g., Williams 1962). A third theory was put forward in the 1970s by the postrevisionist school, which saw the Cold War as primarily the product of misunderstandings that could have been avoided at least to some degree (e.g., Gaddis 1972). If we are interested in testing a liberal theory dealing with the impact of perceptions and misperceptions, we would find supporting evidence for the theory in the postrevisionist school and disconfirming evidence in the previous two schools. In Bayesian terms, this form of bias implies a form of rigging of test results that undermines our ability to update our confidence in the accuracy of hypothesis.

Lustick's solution to this bias problem is first to know the historiographic schools and then to triangulate across them. The researcher should select the historical work that is best suited to provide a critical test. If one is choosing to test an ideational mechanism dealing with actor perceptions, one should not choose work from the postrevisionist school as a source of evidence.

Newspaper Sources

Finally, newspaper and other journalistic sources can, in certain circumstances, provide accurate observations of what we intend to measure. In the case of Owen's tests of part 5 of his hypothesized democratic peace mechanism (see chapter 5), editorials in the liberal press about views toward France

were a form of primary evidence of "liberal elites agitating." Newspapers can provide pattern evidence (e.g., number of articles dealing with a subject), sequence evidence (timetable of events), and different forms of account evidence.

However, the accuracy of newspaper sources often can be difficult to assess. In the words of Moravcsik, "Journalists generally repeat the justifications of governments or the conventional wisdom of the moment without providing much with which to judge the nature or reliability of the source. Second and more important, even . . . if reliable, their sheer number and diversity means that the ability of an analyst to present such evidence tells us little or nothing" (1998: 81). Here we agree with Larson, who suggests that these forms of secondary sources can provide important background material about the context in which decisions were made and what events took place they cannot be used as evidence in process-tracing research unless the observations are triangulated with other types of sources. In other words, newspaper sources are often better at supplying k than they are at supplying o in the formula $o + k \rightarrow e$.