Chapter 3

The Method of Structured, Focused Comparison

The method and logic of structured, focused comparison is simple and straightforward. The method is "structured" in that the researcher writes general questions that reflect the research objective and that these questions are asked of each case under study to guide and standardize data collection, thereby making systematic comparison and cumulation of the findings of the cases possible. The method is "focused" in that it deals only with certain aspects of the historical cases examined. The requirements for structure and focus apply equally to individual cases since they may later be joined by additional cases.

The method was devised to study historical experience in ways that would yield useful generic knowledge of important foreign policy problems. The particular challenge was to analyze phenomena such as deterrence in ways that would draw the explanations of each case of a particular phenomenon into a broader, more complex theory. The aim was to discourage decision-makers from relying on a single historical analogy in dealing with a new case.[1]

---

1. This discussion draws upon earlier publications: Alexander L. George, "Case Studies and Theory Development: The Method of Structured, Focused Comparison," in Paul Gordon Lauren, ed., *Diplomacy: New Approaches in History, Theory, and Policy* (New York: Free Press, 1979), pp. 43–68; Alexander L. George, "The Causal Nexus Between Cognitive Beliefs and Decision-Making Behavior," in Lawrence S. Falkowski, ed., *Psychological Models in International Politics* (Boulder, Colo.: Westview Press, 1979), pp. 95–124; and Alexander L. George and Timothy J. McKeown, "Case Studies and Theories of Organizational Decision Making," in Robert F. Coulam and Richard A. Smith, eds., *Advances in Information Processing in Organizations*, Vol. 2 (Greenwich, Conn.: JAI Press, 1985), pp. 21–58.

An extension of structured, focused comparison is proposed by Patrick J. Haney in

Before we discuss each of these two characteristics of structured, focused comparison, it will be instructive to show how they improve upon previous case study approaches. Following the end of World War II, many political scientists were quite favorably disposed toward or even enthusiastic about the prospect of undertaking individual case studies for the development of knowledge and theory. Many case studies were conducted, not only in the field of international relations but also in public administration, comparative politics, and American politics. Although individual case studies were often instructive, they did not lend themselves readily to strict comparison or to orderly cumulation. As a result, the initial enthusiasm for case studies gradually faded, and the case study as a strategy for theory development fell into disrepute.[2] In 1968 James Rosenau critiqued case studies of foreign policy and called attention to their nonscientific, noncumulative character. These studies of foreign policy by political scientists and historians, Rosenau observed, were not conducted in ways appropriate for scientific inquiry. In his view, most of them lacked "scientific consciousness" and did not accumulate. Individual studies may have made interesting contributions to knowledge, but a basis for systematic comparison was lacking.[3]

---

his *Organizing for Foreign Policy Crisis* (Ann Arbor, Mich.: University of Michigan Press, 1997). Haney develops ways of surveying cases that are capable of combining the advantages of structured, focused comparison with large-N analysis. He suggests that the findings of a number of studies that address the same problem can be combined and the results averaged—i.e., a form of what statisticians refer to as "meta-analysis." This particular case survey method was proposed earlier by Robert Yin and Karen A. Heald, "Using the Case Survey Method to Analyze Policy Studies," *Administrative Science Quarterly*, Vol. 20, No. 3 (September 1975), pp. 371–381. The rather obvious limitations of the case survey approach are noted by Yin and Heald.

A cogent statement of key research steps in small-n research is provided by Ronald Mitchell and Thomas Bernauer in "Empirical Research in International Environmental Policy: Designing Qualitative Case Studies," *Journal of Environment and Development*, Vol. 7, No. 1 (March 1998), pp. 4–31.

Dwaine Medford outlines a way of extending and generalizing structured, focused comparisons that focus on the actor's cognitive processes in Charles F. Hermann, Charles W. Kegley, Jr., and James N. Rosenau, eds., *New Directions in the Study of Foreign Policy* (Boston, Mass.: Allen & Unwin, 1987).

See also our commentary on the important work by Thomas Homer-Dixon in the Appendix, "Studies That Illustrate Research Design."

2. Of course, as noted in Chapter 10, well-researched case studies that are largely descriptive and atheoretical are useful in providing a form of vicarious experience for students and others interested in a particular phenomenon, and sometimes they provide data that can be of some use in case studies devoted to theory development.

3. James N. Rosenau, "Moral Fervor, Systematic Analysis, and Scientific Consciousness in Foreign Policy Research," in Austin Ranney, ed., *Political Science and Public Policy* (Chicago, Ill.: Markham, 1968), pp. 197–238.

Writers in other fields of political science offered similar critiques of extant case studies. In 1955, Roy Macridis and Bernard Brown criticized the old "comparative politics" for being, among other things, not genuinely comparative. These earlier studies consisted mainly of single case studies which were often essentially descriptive and monographic rather than theory-oriented. In the field of public administration, similar concerns were expressed, and, in the field of American politics, an important critique of the atheoretical case study was presented by Theodore Lowi.[4]

What, then, are some of the requirements that case study research must meet to overcome these difficulties?

First, the investigator should clearly identify the universe—that is, the "class" or "subclass" of events—of which a single case or a group of cases to be studied are instances. Thus, the cases in a given study must all be instances, for example, of only one phenomenon: either deterrence, coercive diplomacy, crisis management, alliance formation, war termination, the impact of domestic politics on policymaking, the importance of personality on decision-making, or whatever else the investigator wishes to study and theorize about. The identification of the class or subclass of events for any given study depends upon the problem chosen for study.

Second, a well-defined research objective and an appropriate research strategy to achieve that objective should guide the selection and analysis of a single case or several cases within the class or subclass of the phenomenon under investigation. Cases should not be chosen simply because they are "interesting" or because ample data exist for studying them.

Third, case studies should employ variables of theoretical interest for purposes of explanation. These should include variables that provide some leverage for policymakers to enable them to influence outcomes.

We turn now to a discussion of the two characteristics of the method of structured, focused comparison. From the statistical (and survey) research model, the method of structured, focused comparison borrows the device of asking a set of standardized, general questions of each case, even in single case studies. These questions must be carefully developed to reflect the research objective and theoretical focus of the inquiry. The use of a set of general questions is necessary to ensure the acquisition of comparable data in comparative studies. This procedure allows researchers to avoid the all too familiar and disappointing pitfall of traditional, in-

---

4. Roy C. Macridis and Bernard E. Brown, eds., *Comparative Politics: Notes and Readings* (Homewood, Ill.: Dorsey Press, 1955); Herbert Kaufmann, "The Next Step in Case Studies," *Public Administration Review*, Vol. 18 (Winter 1958), pp. 52–59; and Theodore J. Lowi, "American Business, Public Policy, Case-Studies and Political Theory," *World Politics*, Vol. 16, No. 1 (July 1964), pp. 671–715.

tensive single case studies. Even when such cases were instances of a class of events, they were not performed in a comparable manner and hence did not contribute to an orderly, cumulative development of knowledge and theory about the phenomenon in question. Instead, each case study tended to go its own way, reflecting the special interests of each investigator and often being unduly shaped by whatever historical data was readily available. As a result, idiosyncratic features of each case or the specific interests of each investigator tended to shape the research questions. Not surprisingly, single case studies—lacking "scientific consciousness"—did not accumulate.

The method also requires that the study of cases be "focused": that is, they should be undertaken with a specific research objective in mind and a theoretical focus appropriate for that objective. A single study cannot address all the interesting aspects of a historical event. It is important to recognize that a single event can be relevant for research on a variety of theoretical topics. For example, the Cuban Missile Crisis offers useful material for developing many different theories. This case may be (indeed, has been) regarded and used as an instance of deterrence, coercive diplomacy, crisis management, negotiation, domestic influence on foreign policy, personality involvement in decision-making, etc. Each of these diverse theoretical interests requires the researcher to adopt a different focus, to develop and use a different theoretical framework, and to identify a different set of data requirements. A researcher's treatment of a historical episode must be selectively focused in accordance with the type of theory that the investigator is attempting to develop.

One reason so many case studies of a particular phenomenon in the past did not contribute much to theory development is that they lacked a clearly defined and common focus. Different investigators engaged in research on a particular phenomenon tended to bring diverse theoretical (and nontheoretical) interests to bear on their case studies. Each case study tended to investigate somewhat different dependent and independent variables. Moreover, many case studies were not guided by a well-defined theoretical objective. Not surprisingly, later researchers who had a well-defined theoretical interest in the phenomenon often found that earlier studies were of little value for their purposes.

It is important for researchers to build self-consciously upon previous studies and variable definitions as much as possible—including studies using formal, statistical, and qualitative methods. "Situating" one's research in the context of the literature is key to identifying the contribution the new research makes. Of course, researchers will sometimes find it necessary to modify existing definitions of variables or add new ones, but they must be precise and clear in doing so and acknowl-

edge that this reduces the comparability to or cumulativity with previous studies.

It should be noted that a merely formalistic adherence to the format of structured, focused comparison will not yield good results. The important device of formulating a set of standardized, general questions to ask of each case will be of value only if those questions are grounded in—and adequately reflect—the theoretical perspective and research objectives of the study. Similarly, a selective theoretical focus for the study will be inadequate by itself unless coupled with a relevant set of standardized general questions.

In comparative case studies, structure and focus are easier to achieve if a single investigator not only plans the study, but also conducts all of the case studies. Structured, focused comparison is more difficult to carry out in collaborative research when each case study is undertaken by a different scholar. Collaborative studies must be carefully planned to impress upon all participants the requirements of structure and focus. The chief investigator must monitor the conduct of case studies to ensure that the guidelines are observed by the case writers and to undertake corrective actions if necessary. Properly coordinating the work of case writers in a collaborative study can be a challenging task for the chief investigator, particularly when the contributors are well-established scholars with views of their own regarding the significance of the case they are preparing.

This can be seen in comparing two collaborative studies. One study of Western democratic political opposition brought together a distinguished group of scholars, each studying the democratic opposition in a Western democracy. The study was not tightly organized to meet the requirement of a structured comparison, so the organizer of the study was left with the difficult task of drawing together the disparate findings of the individual case studies for comparative analysis in the concluding chapter.[5] In contrast, Michael Krepon and Dan Caldwell developed a tight version of structured, focused comparison for their collaborative study of cases of U.S. Senate ratification of arms control treaties. They

---

5. Robert A. Dahl, *Political Oppositions in Western Democracies* (New Haven, Conn.: Yale University Press, 1966). As Sidney Verba notes in his detailed commentary on this book, it "highlights a problem that arises in the multiauthored book. There are great advantages in having a large number of country specialists, but specialists are hard to discipline. In *Political Oppositions*, the major theoretical chapters that attempt to tie together the individual country chapters are found at the end of the book. . . . If we want to have as collaborators men of the stature of the authors of this book, we must let them go their own way." Sidney Verba, "Some Dilemmas in Comparative Research," *World Politics*, Vol. 20, No. 1 (October 1967), pp. 116–118).

closely monitored the individual authors' adherence to the guidelines and intervened as necessary to ensure that they adhered to the original or revised guidelines.[6]

The next chapter provides a more specific discussion of procedures for the design and implementation of case studies—either single case analyses or comparative investigations that are undertaken within the framework of the structured, focused method.

---

6. Michael Krepon and Dan Caldwell, eds., *The Politics of Arms Control Treaty Ratification* (New York: St. Martin's Press, 1991). We are indebted to Michael Krepon for providing us with a detailed account of how he and Caldwell accomplished this difficult task.

Chapter 4

Phase One: Designing Case Study Research

There are three phases in the design and implementation of theory-oriented case studies. In phase one, the objectives, design, and structure of the research are formulated. In phase two, each case study is carried out in accordance with the design. In phase three, the researcher draws upon the findings of the case studies and assesses their contribution to achieve the research objective of the study. These three phases are interdependent, and some iteration is often necessary to ensure that each phase is consistent and integrated with the other phases.[1] The first phase is discussed in this chapter, and phases two and three in the chapters that follow.

Phase one—the research design—consists of five tasks. These tasks are relevant not only for case study methodology but for all types of systematic, theory-oriented research. They must be adapted, of course, to different types of investigation and to whether theory testing or theory development is the focus of the study. The design phase of theory-oriented case study research is of critical importance. If a research design

---

1. The procedure of organizing such studies on the basis of these three phases was introduced by Alexander L. George and Richard Smoke in their book *Deterrence in American Foreign Policy: Theory and Practice* (New York: Columbia University Press, 1974). It has proven to be a useful organizing device in subsequent studies and has also provided a framework for reviewing and evaluating existing studies. We are omitting here a fourth phase, presentation of the results of the study, that was mentioned in Alexander L. George and Timothy J. McKeown, "Case Studies and Theories of Organizational Decision Making," in Robert F. Coulam and Richard A. Smith, eds., *Advances in Information Processing in Organizations*, Vol. 2 (Greenwich, Conn.: JAI Press, 1985), pp. 21–58. Some of the observations therein are discussed in the treatment of Phase Two in the present study.

proves inadequate, it will be difficult to achieve the research objectives of the study. (Of course, the quality of the study depends also on how well phases two and three are conducted.)

Task One: Specification of the Problem and Research Objective

The formulation of the research objective is the most important decision in designing research. It constrains and guides decisions that will be made regarding the other four tasks.

The selection of one or more objectives for research is closely coupled with identification of an important research problem or "puzzle." A clear, well-reasoned statement of the research problem will generate and focus the investigation. A statement that merely asserts that "the problem is important" is inadequate. The problem should be embedded in a well-informed assessment that identifies gaps in the current state of knowledge, acknowledges contradictory theories, and notes inadequacies in the evidence for existing theories. In brief, the investigator needs to make the case that the proposed research will make a significant contribution to the field.

The research objective must be adapted to the needs of the research program at its current stage of development. Is there a need for testing a well-established theory or competing theories? Is it important to identify the limits of a theory's scope? Does the state of research on the phenomenon require incorporation of new variables, new subtypes, or work on different levels of analysis? Is it considered desirable at the present stage of theory development to move up or down the ladder of generality?[2] For example, as noted in Chapter 2, in the 1990s the democratic peace research program moved largely from the question of whether such a peace existed to that of identifying the basis on which democratic peace rests. It now needs to go further to explain how a particular peace between two democratic states developed over time. Similarly, in the 1960s deterrence theory needed to bring in additional variables to add to excessively parsimonious and abstract deductive models.

In general, there are six different kinds of theory-building research objectives. Arend Lijphart and Harry Eckstein identified five types. We outline these below and add a sixth type of our own:[3]

---

2. Giovanni Sartori, "Concept Misformation in Comparative Politics," *American Political Science Review,* Vol. 64, No. 4 (December 1970), pp. 1033–1053.

3. Arend Lijphart, "Comparative Politics and the Comparative Method," *American Political Science Review,* Vol. 65, No. 3 (September 1971), pp. 682–693; and Harry Eckstein, "Case Studies and Theory in Political Science," in Fred Greenstein and Nel-

- *Atheoretical/configurative idiographic* case studies provide good descriptions that might be used in subsequent studies for theory building, but by themselves, such cases do not cumulate or contribute directly to theory.
- *Disciplined configurative* case studies use established theories to explain a case. The emphasis may be on explaining a historically important case, or a study may use a case to exemplify a theory for pedagogical purposes. A disciplined configurative case can contribute to theory testing because it can "impugn established theories if the theories ought to fit it but do not," and it can serve heuristic purposes by highlighting the "need for new theory in neglected areas."[4] However, a number of important methodological questions arise in using disciplined configurative case studies and these are discussed in Chapter 9 on the congruence method.
- *Heuristic* case studies inductively identify new variables, hypotheses, causal mechanisms, and causal paths. "Deviant" or "outlier" cases may be particularly useful for heuristic purposes, as by definition their outcomes are not what traditional theories would anticipate. Also, cases where variables co-vary as expected but are at extremely high or low values may help uncover causal mechanisms.[5] Such cases may not allow inferences to wider populations if relationships are nonlinear or involve threshold effects, but limited inferences might be possible if causal mechanisms are identified (just as cancer researchers use high dosages of potential carcinogens to study their effects).
- *Theory testing* case studies assess the validity and scope conditions of single or competing theories. As discussed in Chapter 6, it is important in tests of theories to identify whether the test cases are most-likely, least-likely, or crucial for one or more theories. Testing may also be devised to identify the scope conditions of theories (the conditions under which they are most- and least-likely to apply).
- *Plausibility probes* are preliminary studies on relatively untested theories and hypotheses to determine whether more intensive and laborious testing is warranted. The term "plausibility probe" should not be used too loosely, as it is not intended to lower the standards of evidence and inference and allow for easy tests on most-likely cases.

---

son Polsby, eds., *Handbook of Political Science,* Vol. 7 (Reading, Mass.: Addison-Wesley, 1975), pp. 79–138.

4. Eckstein, "Case Studies and Theory," p. 99.

5. Stephen Van Evera, *Guide to Methods for Students of Political Science* (Ithaca, N.Y.: Cornell University Press, 1997).

- *"Building Block" studies of particular types or subtypes* of a phenomenon identify common patterns or serve a particular kind of heuristic purpose. These studies can be component parts of larger contingent generalizations and typological theories. Some methodologists have criticized single-case studies and studies of cases that do not vary in their dependent variable.[6] However, we argue that single-case studies and "no variance" studies of multiple cases can be useful if they pose "tough tests" for theories or identify alternative causal paths to similar outcomes when equifinality is present.[7] (See also the more detailed discussion of "building blocks" theory below.)

Researchers should clearly identify which of these six types of theory-building is being undertaken in a given study; readers should not be left to find an answer to this question on their own. The researcher may fail to make it clear, for example, whether the study is an effort at theory testing or merely a plausibility probe. Or the researcher may fail to indicate whether and what kind of "tough test" of the theory is supposedly being conducted.[8]

These six research objectives vary in their uses of induction and deduction. Also, a single research design may be able to accomplish more than one purpose—such as heuristic and theory testing goals—as long as it is careful in using evidence and making inferences in ways appropriate to each research objective. For example, while it is not legitimate to derive a theory from a set of data and then claim to test it on the same data, it is sometimes possible to test a theory on different data, or new or previously unobserved facts, from the same case.[9]

---

6. Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton, N.J.: Princeton University Press, 1994).

7. David Collier, "Translating Quantitative Methods for Qualitative Researchers: The Case of Selection Bias," *American Political Science Review,* Vol. 89, No. 2 (June 1995), pp. 461–466; and Ronald Rogowski, "The Role of Theory and Anomaly in Social-Science Inference," *American Political Science Review,* Vol. 89, No. 2 (June 1995), pp. 467–470. Theory development via building blocks is useful also in the absence of equifinality. Contingent generalizations are possible, and indeed easier to formulate, when equifinality is not present. For an example of this approach see George and Smoke, *Deterrence in American Foreign Policy.*

8. Joseph Grieco criticizes Robert O. Keohane's *After Hegemony: Cooperation and Discord in the World Political Economy* (Princeton, N.J.: Princeton University Press, 1984) on both counts in his detailed criticism of the research design in this important study, to which Keohane replies in David A. Baldwin, ed., *Neorealism and Neoliberalism: The Contemporary Debate* (New York: Columbia University Press, 1993).

9. Van Evera, *Guide to Methods.*

Specific questions that need to be addressed in designating the research objectives include:

- What is the phenomenon or type of behavior that is being singled out for examination; that is, what is the class or subclass of events of which the cases will be instances?
- Is the phenomenon to be explained thought to be an empirical universal (i.e., no variation in the dependent variable), so that the research problem is to account for the lack of variation in the outcomes of the cases? Or is the goal to explain an observable variation in the dependent variable?
- What theoretical framework will be employed? Is there an existing theory or rival candidate theories that bear on those aspects of the phenomenon or behavior that are to be explained? If not, what provisional theory or theories will the researcher formulate for the purpose of the study? If provisional theories are lacking, what theory-relevant variables will be considered?
- Which aspects of the existing theory or theories will be singled out for testing, refinement, or elaboration?
- If the research objective is to assess the causal effects or the predictions of a particular theory (or independent variable), is that theory sufficiently specified and operationalized to enable it to make specific predictions, or is it only capable of making probabilistic or indeterminate predictions? What other variables and/or conditions need to be taken into account in assessing its causal effects?

Researchers' initial efforts to formulate research objectives for a study often lack sufficient clarity or are too ambitious. Unless these defects are corrected, the study will lack a clear focus, and it will probably not be possible to design a study to achieve the objectives.

Better results are achieved if the "class" of the phenomenon to be investigated is not defined too broadly. Most successful studies, in fact, have worked with a well-defined, smaller-scope *subclass* of the general phenomenon.[10] Case study researchers often move down the "ladder of generality" to contingent generalizations and the identification of more circumscribed scope conditions of a theory, rather than up toward broader but less precise generalizations.[11]

---

10. For illustrative examples, see the Appendix, "Studies That Illustrate Research Design."

11. A similar point is made by Robert Keohane in his critique of structural realism. He notes that it is desirable to select a smaller subclass of a phenomenon in order "to achieve greater precision" of a theory. This entails "narrowing" the "domain of a the-

Working with a specified subclass of a general phenomenon is also an effective strategy for theory development. Instead of trying in one study to develop a general theory for an entire phenomenon (e.g., all "military interventions"), the investigator should think instead of formulating a typology of different kinds of interventions and proceed to choose one type or subclass of interventions for study, such as "protracted interventions." Or the study may focus on interventions by various policy instruments, interventions on behalf of different goals, or interventions in the context of different alliance structures or balances of power. The result of any single circumscribed study will be one part of an overall theory of intervention. Other studies, focusing on different types or subclasses of intervention, will be needed to contribute to the formulation of a general theory of interventions, if that is the broader, more ambitious research program. If the typology of interventions identifies six major kinds of intervention that are deemed to be of theoretical and practical interest, each subtype can be regarded as a candidate for separate study and each study will investigate instances of that subtype.

This approach to theory development is a "building block" procedure. Each block—a study of each subtype—fills a "space" in the overall theory or in a typological theory. In addition, the component provided by each building block is itself a contribution to theory; though its scope is limited, it addresses the important problem or puzzle associated with the type of intervention that led to the selection and formulation of the research objective. Its generalizations are more narrow and contingent than those of the general "covering laws" variety that some hold up as the ideal, but they are also more precise and may involve relations with higher probabilities.[12] In other words, the building block developed for a subtype is self sufficient; its validity and usefulness do not depend upon the existence of other studies of different subclasses of that general phenomenon.

If an investigator wishes to compare and contrast two or more different types of intervention, the study must be guided by clearly defined puzzles, questions, or problems that may be different from or similar to those of a study of a single subclass. For example, the objective may be to discover under what conditions (and through what paths) Outcome X occurs, and under what conditions (and through what paths) Outcome Y

---

ory." Robert O. Keohane, ed., *Neorealism and Its Critics* (New York: Columbia University Press, 1986), pp. 187–188.

12. For example, see the discussion in the Appendix of Ariel Levite, Bruce Jentleson, and Larry Berman, eds., *Foreign Military Intervention: The Dynamics of Protracted Conflict* (New York: Columbia University Press, 1992). See also the discussion of "middle-range" theory in Chapter 12.

occurs. Alternatively, the objective may be to examine under what conditions Policy A leads to Outcome Y and under what other conditions Policy A leads to Outcome X. Similarly, the focus may be on explaining the outcome of a case or a subclass or type of cases, or it may be on explaining the causal role of a particular independent variable across cases.

Task Two: Developing a Research Strategy: Specification of Variables

In the course of formulating a research objective for the study—which may change during the study—the investigator also develops a *research strategy* for achieving that objective. This requires early formulation of hypotheses and consideration of the elements (conditions, parameters, and variables) to be employed in the analysis of historical cases. Several basic decisions (also subject to change during the study) must be made concerning questions such as the following:

- What exactly and precisely is the dependent (or outcome) variable to be explained or predicted?
- What independent (and intervening) variables comprise the theoretical framework of the study?
- Which of these variables will be held constant (serve as parameters) and which will vary across cases included in the comparison?

The specification of the problem in Task One is closely related to the statement of what exactly the dependent variable will be. If a researcher defines the problem too broadly, he or she risks losing important differences among cases being compared. If a researcher defines the problem too narrowly, this may severely limit the scope and relevance of the study and the comparability of the case findings.[13] As will be noted, the definition of variance in the dependent variable is critical in research design.

In analyzing the phenomenon of "war termination," for instance, a researcher would specify numerous variables. The investigator would decide whether the dependent (outcome) variable to be explained (or predicted) was merely a cease-fire or a settlement of outstanding issues over which the war had been fought. Variables to be considered in explaining the success or failure of war termination might include the fighting capabilities and morale of the armed forces, the availability of

13. This research dilemma is discussed by Sidney Verba in his detailed commentary on Robert A. Dahl, ed., *Political Oppositions in Western Democracies* (New Haven, Conn.: Yale University Press, 1966), and in Sidney Verba, "Some Dilemmas in Comparative Research," *World Politics*, Vol. 20, No. 1 (October 1967), pp. 122–123.

economic resources for continuing the war, the type and magnitude of pressures from more powerful allies, policymakers' expectation that the original war aim was no longer attainable at all or only at excessive cost, the pressures of pro-war and anti-war opinion at home, and so on. The researcher might choose to focus on the outcome of the dependent variable (e.g., on cases in which efforts to achieve a cease-fire or settlement failed, but adding cases of successful cease-fires or settlements for contrast) to better identify the independent and intervening variables associated with such failures. Alternatively, one might vary the outcome, choosing cases of both successes and failures in order to identify the conditions and variables that seem to account for differences in outcomes.

Alternatively, the research objective may focus not on outcomes of the dependent variable, but on the importance of an independent variable—e.g., war weariness—in shaping outcomes in a number of cases.

We conclude this discussion of Task Two with a brief review of the strengths and weaknesses of the common types of case study research designs in relation to the kinds of research objectives noted above.

First, single case research designs can fall prey to selection bias or over-generalization of results, but all of the six theory-building purposes identified above have been served by studies of single well-selected cases that have avoided or minimized such pitfalls. Obviously, single-case studies rely almost exclusively on within-case methods, process-tracing, and congruence, but they may also make use of counterfactual analysis to posit a control case.[14]

For theory testing in single cases, it is imperative that the process-tracing procedure and congruence tests be applied to a wide range of alternative hypotheses that theorists and even participants in the events have proposed, not only to the main hypotheses of greatest interest to the researcher. Otherwise, *left-out variables* may threaten the validity of the research design. Single cases serve the purpose of theory testing particularly well if they are "most-likely," "least-likely," or "crucial" cases. Prominent case studies by Arend Lijphart, William Allen, and Peter Gourevitch, for example, have changed entire research programs by impugning theories that failed to explain their most-likely cases.[15]

---

14. David Laitin, "Disciplining Political Science," *American Political Science Review*, Vol. 89, No. 2 (June 1995), pp. 454–456. We say "almost" since single case studies take place within the context of ongoing research programs, so that studies of single cases may draw comparisons to existing studies; thus, "the community of scientists," rather than the "individual researcher" is the relevant context in which to judge case selection.

15. Rogowski, "The Role of Theory and Anomaly in Social-Scientific Inference"; Arend Lijphart, *The Politics of Accommodation: Pluralism and Democracy in the Netherlands* (Berkeley: University of California Press, 1968); William Sheridan Allen, *The Nazi*

Similarly, studies of single "deviant" cases and of single cases where a variable is at an extreme value can be very useful for heuristic purposes of identifying new theoretical variables or postulating new causal mechanisms. Single-case studies can also serve to reject variables as being necessary or sufficient conditions.[16]

Second, the research objective chosen in a study may require comparison of several cases. There are several comparative research designs. The best known is the method of "controlled comparison"—i.e., the comparison of "most similar" cases which, ideally, are cases that are comparable in all respects except for the independent variable, whose variance may account for the cases having different outcomes on the dependent variable. In other words, such cases occupy neighboring cells in a typology, but only if the typological space is laid out one change in the independent variable at a time. (See Chapter 11 on typological theories.)

As we discuss in Chapter 8 on the comparative method, controlled comparison can be achieved by dividing a single longitudinal case into two—the "before" case and an "after" case that follows a discontinuous change in an important variable. This may provide a control for many factors and is often the most readily available or strongest version of a most-similar case design. This design aims to isolate the difference in the observed outcomes as due to the influence of variance in the single independent variable. Such an inference is weak, however, if the posited causal mechanisms are probabilistic, if significant variables are left out of the comparison, or if other important variables change in value from the "before" to the "after" cases.

However, even when two cases or before-after cases are not perfectly matched, process-tracing can strengthen the comparison by helping to assess whether differences other than those in the main variable of interest might account for the differences in outcomes. Such process-tracing can focus on the standard list of potentially "confounding" variables identified by Donald Campbell and Julian Stanley, including the effects of history, maturation, testing, instrumentation, regression, selection, and mortality.[17] It can also address any idiosyncratic differences between the two

---

Seizure of Power: The Experience of a Single German Town, 1930–1935 (New York: Watts, 1965); and Peter Alexis Gourevitch, "The International System and Regime Formation: A Critical Review of Anderson and Wallerstein," *Comparative Politics*, Vol. 10, No. 3 (April 1978), pp. 419–438.

16. For an example, see Lijphart's study summarized in the Appendix, "Studies That Illustrate Research Design"; Douglas Dion, "Evidence and Inference in Comparative Case Study," *Comparative Politics*, Vol. 3, No. 2 (January 1998); and Collier, "Translating Quantitative Methods," p. 464.

17. Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental*

cases that scholars or participants have argued might account for their differences.

Another comparative design involves "least similar" cases and parallels John Stuart Mill's method of agreement.[18] Here, two cases are similar in outcome but differ in all but one independent variable, and the inference might be made that this variable contributes to the invariant outcome. For example, if teenagers are "difficult" in both postindustrial societies and tribal societies, we might infer that their developmental stage, and not their societies or their parents' child-rearing techniques, account for their difficult natures. Here again, left-out variables can weaken such an inference, as Mill recognized, but process-tracing provides an additional source of evidence for affirming or infirming such inferences.

Another type of comparative study may focus on cases in the same cell of a typology. If these have the same outcome, process-tracing may still reveal different causal paths to that outcome. Conversely, multiple studies of cases with the same level of a manipulable independent variable can establish under what conditions that level of the variable is associated with different outcomes. In either approach, if outcomes differ within the same type or cell, it is necessary to look for left-out variables and perhaps create a new subtype.

Often, it is useful for a community of researchers to study or try to identify cases in all quadrants of a typology. For example, Sherlock Holmes once inferred that a dog that did not bark must have known the person who entered the dog's house and committed a murder, an inference based on a comparison to dogs that do bark in such circumstances. To fully test such an assertion, we might also want to consider the behavior of non-barking non-dogs on the premises (was there a frightened cat?) and barking non-dogs (such as a parrot). The process of looking at all the types in a typology corresponds with notions of Boolean algebra and those of logical truth tables.[19] However, it is not necessary for each researcher to address all the cells in a typology, although it is often useful

---

Designs for Research (Chicago: Rand McNally College Publishing, 1963); for a good example, see James Lee Ray, *Democracies and International Conflict: An Evaluation of the Democratic Peace Proposition* (Columbia: University of South Carolina Press, 1995), pp. 158–200.

18. For a detailed discussion of Mill's methods, see Chapter 8.

19. Charles C. Ragin, *The Comparative Method* (Berkeley: University of California Press, 1987); Daniel Little, *Varieties of Social Explanation: An Introduction to the Philosophy of Science* (Boulder, Colo.: Westview Press, 1991); and Daniel Little, *Microfoundations, Method, and Causation: On the Philosophy of the Social Sciences* (New Brunswick, N.J.: Transaction Publishers, 1998).

for researchers to offer suggestions for future research on unexamined types or to make comparisons to previously examined types.

Finally, a study that includes many cases may allow for several different types of comparisons. One case may be most similar to another and both may be least similar to a third case. As noted below, case selection is an opportunistic as well as a structured process—researchers should look for whether the addition of one or a few cases to a study might provide useful comparisons or allow inferences on additional types of cases.

Task Three: Case Selection

Many students in the early stages of designing a study indicate that they find it difficult to decide which cases to select. This difficulty usually arises from a failure to specify a research objective that is clearly formulated and not overly ambitious. One should select cases not simply because they are interesting, important, or easily researched using readily available data. Rather, case selection should be an integral part of a good research strategy to achieve well-defined objectives of the study. Hence, the primary criterion for case selection should be relevance to the research objective of the study, whether it includes theory development, theory testing, or heuristic purposes.

Cases should also be selected to provide the kind of control and variation required by the research problem. This requires that the universe or subclass of events be clearly defined so that appropriate cases can be selected. In one type of comparative study, for example, all the cases must be instances of the same subclass. In another type of comparative study that has a different research objective, cases from different subclasses are needed.

Selection of a historical case or cases may be guided by a typology developed from the work in Tasks One and Two. Researchers can be somewhat opportunistic here—they may come across a pair of well-matched before-after cases or a pair of cases that closely fit "most similar" or "least similar" case research designs. They may also come upon cases that have many features of a most- or least-likely case, a crucial case, or a deviant case.

Often researchers begin their inquiry with a theory in search of a test case or a case in search of a theory for which it is a good test.[20] Either approach is viable, provided that care is taken to prevent case selection bias and, if necessary, to study several cases that pose appropriate tests for a

20. King, Keohane, and Verba, *Designing Social Inquiry*, pp. 17–18.

candidate theory once one is identified. Often, the researcher might start with a case that interests her, be drawn to a candidate theory, and then decide that she is more interested in the theory than in the case and conclude that the best way to study the theory is to select several cases that may not include the case with which the inquiry began. Some such iteration is usually necessary—history may not provide the ideal kind of cases to carry out the tests or heuristic studies that a research program most needs at its current stage of development.

Important criticisms have been made of potential flaws in case selection in studies with one or a few cases; such concerns are influenced by the rich experience of statistical methods for analyzing a large-N. David Collier and James Mahoney have taken issue with some widespread concerns about selection bias in small studies; we note four of their observations.[21] They question the assertion that selection bias in case studies is potentially an even greater problem than is often assumed (that it may not just understate relationships—the standard statistical problem—but may overstate them). They argue that case study designs with no variance in the dependent variable do not inherently represent a selection bias problem. They emphasize that case study researchers sometimes have good reasons to narrow the range of cases studied, particularly to capture heterogeneous causal relations, even if this increases the risk of selection bias. They point out (as have we) that case study researchers rarely "overgeneralize" from their cases; instead, they are frequently careful in providing circumscribed "contingent generalizations" that subsequent researchers should not mistakenly overgeneralize.

Task Four: Describing the Variance in Variables

The way in which variance is described is critical to the usefulness of case analyses in furthering the development of new theories or the assessment or refinement of existing theories. This point needs emphasis because it is often overlooked in designing studies—particularly statistical studies of a large-N. The researcher's decision about how to describe variance is important for achieving research objectives because the discovery of potential causal relationships may depend on how the variance in these variables is postulated. Basing this decision on *a priori* judgments may be risky and unproductive; the investigator is more likely to develop sensitive ways of describing variance in the variables after he or she has become familiar with how they vary in the historical cases examined. An it-

21. David Collier and James Mahoney, "Insights and Pitfalls: Selection Bias in Qualitative Research," *World Politics*, Vol. 49, No. 1 (October 1996), pp. 56–91.

erative procedure for determining how best to describe variance is therefore recommended.[22]

The variance may in some instances be best described in terms of qualitative types of outcomes. In others, it may be best described in terms of quantitative measures. In either case, one important question is how many categories to establish for the variables. Fewer categories—such as dichotomous variables—are good for parsimony but may lack richness and nuance, while greater numbers of categories gain richness but sacrifice parsimony. The trade-off between parsimony and extreme richness should be determined by considering the purposes of each individual study.

In a study of deterrence, for example, Alexander George and Richard Smoke found it to be inadequate and unproductive to define deterrence outcomes simply as "successes" or "failures."[23] Instead, their explanations of individual cases of failure enabled them to identify different types of failures. This led to a typology of failures, with each type of failure having a different explanation. This typology allowed George and Smoke to see that deterrence failures exemplified the phenomenon of equifinality. The result was a more discriminating and policy-relevant explanatory theory for deterrence failures.[24]

The differentiation of types can apply to the characterization of independent as well as dependent variables. In attempting to identify conditions associated with the success or failure of efforts to employ a strategy of coercive diplomacy, one set of investigators identified important variants of that strategy.[25] In their study, coercive diplomacy was treated as an independent variable. From an analysis of different cases, four types of the coercive diplomacy strategy were identified: the explicit ultimatum, the tacit ultimatum, the "gradual turning of the screw," and the "try and see" variant. By differentiating the independent variable in this way, it was possible to develop a more discriminating analysis of the effectiveness of coercive diplomacy and to identify some of the factors that favored or handicapped the success of each variant. A very general or undifferentiated depiction of the independent variable would have

---

22. See also the discussion of this point in Chapter 9 on "The Congruence Method."

23. George and Smoke, *Deterrence in American Foreign Policy.*

24. See the Appendix, "Studies That Illustrate Research Design," for a fuller discussion of their study.

25. Alexander L. George, David K. Hall, and William E. Simons, *The Limits of Coercive Diplomacy* (Boston: Little, Brown, 1971); an extended second edition under the same title that examines additional cases was published in 1994, edited by Alexander L. George and William E. Simons (Boulder, Colo.: Westview Press).

"washed out" the fact that variants of coercive diplomacy may have different impacts on outcomes, or it might have resulted in ambiguous or invalid results. In addition, the identification of different variants of coercive diplomacy strategy has important implications for the selection of cases.

Task Five: Formulation of Data Requirements and General Questions

The case study method will be more effective if the research design includes a specification of the data to be obtained from the case or cases under study. Data requirements should be determined by the theoretical framework and the research strategy to be used for achieving the study's research objectives. The specification of data requirements should be integrated with the other four design tasks. Specification of data requirements structures the study. It is an essential component of the method of structured, focused comparison.

Whether a single-case study or a case comparison is undertaken, specification of the data requirements should take the form of general questions to be asked of each case. This is a way of standardizing data requirements so that comparable data will be obtained from each case and so that a single-case study can be compared later with others. Case study methodology is no different in this respect from large-N statistical studies and public opinion surveys. Unless one asks the same questions of each case, the results cannot be compared, cumulated, and systematically analyzed.

This is only to say—and to insist—that case researchers should follow a procedure of systematic data compilation. The questions asked of each case must be of a general nature; they should not be couched in overly specific terms that are relevant to only one case but should be applicable to all cases within the class or subclass of events with which the study is concerned. Asking the same questions of each case does *not* prevent the case writer from addressing more specific aspects of the case or bringing out idiosyncratic features of each case that may also be of interest for theory development or future research.

A problem sometimes encountered in case study research is that data requirements are missing altogether or inadequately formulated. The general questions must reflect the theoretical framework employed, the data that will be needed to satisfy the research objective of the study, and the kind of contribution to theory that the researcher intends to make. In other words, a mechanical use of the method of structured, focused comparison will not yield good results. The proper focusing and structuring of the comparison requires a fine-tuned set of general questions that are

integrated with the four other elements of the research design. For example, in a comparative study of policymakers' approaches to strategy and tactics toward political opponents in the international arena, one might start by asking questions designed to illuminate the orientations of a leader toward the fundamental issues of history and politics that presumably influence his or her processing of information, policy preference, and final choice of action.[26] In this type of study, the investigator examines an appropriate body of material in order to infer the "answers" a political leader might have given to the following questions:

PHILOSOPHICAL QUESTIONS

- What is the essential nature of political life? Is the political universe essentially one of harmony or conflict? What is the fundamental character of one's political opponents?
- What are the prospects for eventual realization of one's fundamental political values and ideological goals? Can one be optimistic or pessimistic?
- In what sense and to what extent is the political future predictable?
- How much control or mastery can one have over historical developments? What is the political leader's (or elite's) role in moving and shaping history?
- What is the role of chance in human affairs and in historical development?

INSTRUMENTAL QUESTIONS

- What is the best approach for selecting goals or objectives for political action?
- How are the goals of action pursued most effectively?
- How are the risks of political action best calculated, controlled, and accepted?
- What is the best timing of action to advance one's interests?

---

26. See Alexander L. George, "The 'Operational Code': A Neglected Approach to the Study of Political Leaders and Decision-Making," *International Studies Quarterly*, Vol. 13, No. 2 (June 1969), pp. 190–222. The problem of judging the causal role of such beliefs in a policymaker's choice of action was discussed in Alexander L. George, "The Causal Nexus Between Cognitive Beliefs and Decision-Making Behavior: The 'Operational Code' Belief System," in Lawrence S. Falkowski, ed., *Psychological Models In International Relations* (Boulder, Colo.: Westview Press, 1979), pp. 95–124. Since then, numerous studies have been made of the "operational codes" of a variety of leaders using this standardized approach or a slight modification of it. This has facilitated comparison and cumulation of results. See, for example, the publications of Ole R. Holsti and Stephen G. Walker.

- What is the utility and role of different means for advancing one's interests?

Integration of the Five Design Tasks

The five design tasks should be viewed as constituting an integrated whole. The researcher should keep in mind that these tasks are interrelated and interdependent. For example, the way in which Task Two is performed should be consistent with the specification of Task One. Similarly, both the selection of cases in Task Three and the theoretical framework developed in Task Four must be appropriate and serviceable from the standpoint of the determinations made for Tasks One and Two. And finally, the identification of data requirements in Task Five must be guided by the decisions made for Tasks One, Two, and Three.

Yet a satisfactory integration of the five tasks usually cannot be accomplished on the first try. A good design does not come easily. Considerable iteration and respecification of the various tasks may be necessary before a satisfactory research design is achieved. The researcher may need to gain familiarity with the phenomenon in question by undertaking a preliminary examination of a variety of cases before finalizing aspects of the design.

Despite the researcher's best efforts, the formulation of the design is likely to remain imperfect—and this may not be apparent until the investigator is well into phase two or even phase three of the study. If these defects are sufficiently serious, the researcher should consider halting further work and redesigning the study, even if this means that some of the case studies will have to be redone. In drawing conclusions from the study, the researcher (or others who evaluate it) may be able to gain some useful lessons for a better design of a new study of the problem.[27]

---

27. For additional discussion of the critical importance of research design, see the "Pedagogical Note to Parts Two and Three."

Chapter 5

Phase Two: Carrying Out the Case Studies

The fifth task in a research design—the formulation of general questions to ask of each of the cases to be studied in phase two—allows the researcher to analyze each case in a way that will provide "answers" to the general questions.[1] These answers—the product of phase two—then constitute the data for the third phase of research, in which the investigator will use case findings to illuminate the research objectives of the study.

Usually one's first step in studying a case with which one is not already intimately familiar is to gather the most easily accessible academic literature and interview data on the case and its context. This preliminary step of immersing oneself in the case, known as "soaking and poking," often leads to the construction of a chronological narrative that helps both the researcher and subsequent readers understand the basic outlines of the case.[2]

---

1. This chapter draws on earlier publications by Alexander L. George, "Case Studies and Theory Development: The Method of Structured, Focused Comparison," in Paul Gordon Lauren, ed., *Diplomacy: New Approaches in Theory, History, and Policy* (New York: Free Press, 1979), pp. 3–68; Alexander L. George, "The Causal Nexus Between Cognitive Beliefs and Decision-Making Behavior," in Lawrence S. Falkowski, ed., *Psychological Models in International Politics* (Boulder, Colo.: Westview Press, 1979), pp. 95–124; and Alexander L. George and Timothy J. McKeown, "Case Studies and Theories of Organizational Decision Making," in Robert F. Coulam and Richard A. Smith, eds., *Advances in Information Processing in Organizations,* Vol. 2 (Greenwich, Conn.: JAI Press, 1985), pp. 21–58.

2. An interesting example of "soaking and poking" and a description of how it mixes inductive and deductive reasoning is found in Richard F. Fenno's *Homestyle* (Boston: Little, Brown, 1978). As noted in the review of his study in the Appendix, Fenno gives a detailed reconstruction of how his interview questions and research design evolved as he undertook subsequent interviews with members of Congress.

After a period of "soaking and poking," the researcher turns to the task of case study analysis, establishing the values of independent and dependent variables in a case through standard procedures of historical inquiry. (If appropriate, the researcher may be able to quantify and scale variables in some fashion.) The researcher should always articulate the criteria employed for "scoring" the variables so as to provide a basis for inter-coder reliability.

Next, the researcher develops explanations for the outcome of each case. This is a matter of detective work and historical analysis rather than a matter of applying an orthodox quasi-experimental design.[3] Social scientists performing case studies will need to familiarize themselves with the craft of the historian's trade—learning, for the context in which the case is embedded, the special difficulties presented by various kinds of evidence that may be available; using multiple weak inferences rather than single strong inferences to buttress conclusions; developing procedures for searching through large masses of data when the objectives of the search are not easily summarized by a few simple search rules.[4]

This chapter provides advice on these topics. The first three sections focus on the provisional nature of case explanations, and the challenges involved in weighing explanations offered by other researchers who have analyzed a given case, and the task of transforming a descriptive explanation for a case into an explanation that adequately reflects the researcher's theoretical framework. We then turn to issues that researchers encounter when working with a variety of primary and secondary materials. Notable issues with secondary sources include the biases of their authors, and a tendency to overestimate the rationality of the policy-making process while underestimating the complexity and the multitude of interests that may be at play. Scholars face numerous issues in assessing the evidentiary value of primary sources. Finally, we describe some of the tasks faced by those who critically read others' case studies, and urge that researchers make their methods as transparent as possible to the reader.

The Provisional Character of Case Explanations

Case explanations must always be considered to be of a provisional character. Therefore, the theoretical conclusions drawn from case study findings (in phase three) will also be provisional. The explanations pro-

3. For discussion of this point, see George and McKeown, "Case Studies and Theories of Organizational Decision Making," pp. 38–39.

4. The nature and requirements of historical explanations are discussed in Chapter 10.

vided by the case writer may be challenged by other scholars on one or another ground—for example, the original research may have overlooked relevant data or misunderstood its significance, failed to consider an important rival hypothesis, and so forth. If case explanations are later successfully challenged, the researcher will have to reassess the implications for any theory that has been developed or tested. Such a reassessment would also be necessary if new historical data bearing on the cases become available at a later date and lead to a successful challenge of earlier explanations.

In seeking to formulate an explanation for the outcome in each case, the investigator employs the historian's method of causal imputation, which differs from the mode of causal inference in statistical-correlational studies. These causal interpretations gain plausibility if they are consistent with the available data and if they can be supported by relevant generalizations for which a measure of validity can be claimed on the basis of existing studies. The plausibility of an explanation is enhanced to the extent that alternative explanations are considered and found to be less consistent with the data, or less supportable by available generalizations.

An investigator must demonstrate that he or she has seriously considered alternative explanations for the case outcome in order to avoid providing the basis for a suspicion, justified or not, that he or she has "imposed" a favored theory or hypothesis as the explanation. Such a challenge is likely if the reader believes that case selection was biased by the investigator's commitment to a particular theory or hypothesis.[5]

The Problem of Competing Explanations

A familiar challenge that case study methods encounter is to reconcile, if possible, conflicting interpretations of a case or to choose between them. This problem can arise when the investigator provides an explanation that differs from an earlier scholar's but does not adequately demonstrate the superiority of the new interpretation. As Olav Njølstad notes, competing explanations may arise from several sources.[6] There are different types of explanation stemming, for example, from historiographical is-

5. The need to avoid selecting cases that favor a particular theory and that constitute easy rather than tough tests of a theory was emphasized in Chapter 4.

6. This brief discussion draws from the fuller discussion of these problems in Chapter 2, "Case Study Methods and Research on the Interdemocratic Peace," which also provides illustrative materials. See also Olav Njølstad's chapter, "Learning from History? Case Studies and the Limits to Theory-Building," in Nils Petter Gleditsch and Olav Njølstad, eds., *Arms Races: Technological and Political Dynamics* (London: Sage Publications, 1990), pp. 220–245. Njølstad also offers several useful suggestions for dealing with these problems, which are summarized in Chapter 2.

sues such as the relative importance of ideology or historical context. Sometimes competing explanations can be equally consistent with the available historical evidence; this makes it difficult to decide which is the correct explanation or, alternatively, whether both interpretations may be part of the overall explanation—i.e., whether the outcome may be overdetermined. Another possibility is that each of the ostensibly competing explanations in fact addresses different parts of a complex longitudinal development. In such cases, the task of the investigator is to identify different turning points in the causal chain and to sort out which independent variables explain each step in the causal chain—for example, those explaining why a war occurred, those that explain the form of the attack, those that explain its timing, and so on. Still another possibility is that the key variable in one explanation is causal and the proposed causal variable in the other explanation is spurious.

The problem of apparently competing explanations may also arise when the rival interpretations address and attempt to explain different aspects of a case and therefore cannot be reconciled. When this happens, the investigator and readers of the case account should not regard the two interpretations as competing with each other. Another possibility is that the rival explanations emerge because the scholars advancing them have simply disagreed on the "facts" of the case.

In any case, if the data and generalizations available to the investigator do not permit him or her to choose from competing explanations, then both explanations for the case should be retained as equally plausible, and the implications of both for theory development should be considered in phase three of the study.

Transforming Descriptive Explanations Into Analytical Explanations

In addition to developing a specific explanation for each case, the researcher should consider transforming the specific explanation into the concepts and variables of the general theoretical framework specified in Task Two.[7] (In Harry Eckstein's terminology, such research is "disciplined-configurative" rather than "configurative-idiographic.") To transform specific explanations into general theoretical terms, the researcher's theoretical framework must be broad enough to capture the major elements of the historical context. That is, the set of independent and inter-

7. For an early discussion of the practice of transforming a historical explanation into an analytical one see Gabriel Almond et al., *Crisis, Choice, and Change: Historical Studies of Political Development* (Boston: Little, Brown, 1973). This study is among those summarized in the Appendix, "Studies That Illustrate Research Design."

vening variables must be adequate to capture and record the essentials of a causal account of the outcome in the case. The dividing line between what is essential and what is not is whether aspects of a causal process in a given case are expected or found to operate across the entire class of cases under consideration. For example, if some instance of organizational decision-making was decisively affected by the fact that one of the key participants in the decision process caught a cold and was unable to attend an important meeting, this would *not* constitute a basis for revising our theory of organizational decision-making to endogenize the susceptibility of actors to disease. It *would*, however, constitute a basis for a general argument about how outcomes are affected by the presence or absence of important potential participants.

Some historians will object to this procedure for transforming a rich and detailed historical explanation into a more abstract and selective one couched in theoretical concepts, arguing that unique qualities of the explanation inevitably will be lost in the process. This is undoubtedly true: some loss of information and some simplification is inherent in any effort at theory formulation or in theoretically formulated explanations. The critical question, however, is whether the loss of information and the simplification jeopardize the validity of the conclusions drawn from the cases for the theory and the utility of that theory. This question cannot be answered abstractly. The transition from a specific to a more general explanation may indeed lead a researcher to dismiss some of the causal processes at work in the case simply because they are not already captured by the general theory or because the researcher fails to recognize a variable's general significance. To say that avoiding these errors depends on the sensitivity and judgment of the researcher, while true, is not very helpful. One slightly more specific guideline is that researchers seem more susceptible to this error when trying to discern new causal patterns than when attempting to evaluate claims about some causal patterns already hypothesized to be operating in a particular case; and second, that the more fine-tuned and concrete the description of variance, the more readily the analysis will accommodate a more differentiated description of the causal processes at work.[8]

To the extent that the case study method has arisen from the practice of historians, it has tended to follow certain procedures that are not really appropriate for social scientists. One feature of most historians' work is a relative lack of concern with or discussion of methodological issues en-

8. See Chapter 4 for a discussion of Task Four and the critical importance of how variance in the variables is described, our caution against *a priori* decisions on such matters, and the desirability of making such determinations after preliminary analysis of the cases.

countered in the performance of research. We believe that case researchers should explicitly discuss the major research dilemmas the case study researcher faced in the analysis of a case and the justifications for solving those dilemmas in a particular way. Therefore, we recommend that the investigator give some indication of how his or her initial expectations about behavior and initial data-collection rules were revised in the course of the study. This would permit readers to make a more informed analysis of the process by which a case and the conclusions based on the case were reached.

Most historians also rely heavily on chronological narrative as an organizing device for presenting the case study materials. Preserving some elements of the chronology of the case may be indispensable for supporting the theory-oriented analysis, and it may be highly desirable to do so in order to enable readers not already familiar with the history of the case to comprehend the analysis. Striking the right balance between a detailed historical description of the case and development of a theoretically-focused explanation of it is a familiar challenge. Analysts frequently feel it necessary to reduce the length of a case study to avoid overly long accounts that exceed the usual limits for journal articles or even books! The more cases, the more difficult this problem becomes.

There is no easy answer to this dilemma. Still, it has been dealt with in a reasonably effective way by a number of writers. A brief résumé of the case at the beginning of the analysis gives readers the essential facts about the development and outcome of the case. The ensuing write-up can blend additional historical detail with analysis.[9] Presentation of a case need not always include a highly detailed or exclusively chronological narrative. As a theory becomes better developed and as research focuses on more tightly defined targets, there will be less need to present overly long narratives. Moreover, narrative accounts of a case can be supplemented by such devices as decision trees, sketches of the internal analytical structure of the explanation, or even computer programs to display the logic of the actors' decisions or the sequence of internal developments within the case.

Some Challenges in Attempting to Reconstruct Decisions

Scholars who attempt to reconstruct the policymaking process in order to explain important decisions face challenging problems. An important limitation of the analysis presented here is that it is drawn solely from the

9. See, for example, how this task was dealt with in studies such as Alexander L. George and Richard Smoke, *Deterrence in American Foreign Policy: Theory and Practice* (New York: Columbia University Press, 1974).

study of U.S. foreign policy.[10] We discuss first the task of acquiring reliable data on factors that entered into the policy process and evaluating their impact on the decision. Political scientists must often rely upon, or at least make use of, historians' research on the policy in question. Such historical studies can be extremely useful to political scientists, but several cautions should be observed in making use of these studies.

First, researchers should forgo the temptation to rely on a single, seemingly authoritative study of the case at hand by a historian. Such a shortcut overlooks the fact that competent historians who have studied that case often disagree on how best to explain it. As Ian Lustick has argued, "the work of historians is not . . . an unproblematic background narrative from which theoretically neutral data can be elicited for the framing of problems and the testing of theories."[11] Lustick approvingly notes Norman Cantor's argument that a historian's work represents "a picture of 'what happened' that is just as much a function of his or her personal commitments, the contemporary political issues with which s/he was engaged, and the methodological choices governing his or her work."[12] The danger here, Lustick argues, is that a researcher who draws upon too narrow a set of historical accounts that emphasizes the variables of interest may overstate the performance of favored hypotheses.

It is thus necessary to identify and summarize important debates among historians about competing explanations of a case, and wherever possible to indicate the possible political and historical biases of the contending authors. The researcher should translate these debates into the competing hypotheses and their variables as outlined in phase one. If there are important historical interpretations of the case that do not easily translate into the hypotheses already specified, the researcher should consider whether these interpretations should be cast as additional hypotheses and specified in terms of theoretical variables. The same procedures apply to the primary political debates among participants in the case and their critics. Even such overtly political debates may draw upon

---

10. Similar problems arise in efforts by scholars to make use of archival materials and interviews from Soviet sources. See, for example, the correspondence between Mark Kramer, who expressed concern about the use of oral histories by Bruce J. Allyn, James G. Blight, and David A. Welch, and their responses in "Remembering the Cuban Missile Crisis: Should We Swallow Oral History?" *International Security*, Vol. 15, No. 1 (Summer 1990), pp. 212–218. See also "Commentaries on 'An Interview With Sergo Mikoyan'" by Raymond L. Garthoff, Barton J. Bernstein, Marc Trachtenberg, and Thomas G. Paterson in *Diplomatic History*, Vol. 14, No. 2 (Spring 1990), pp. 223–256.

11. Ian S. Lustick, "History, Historiography, and Political Science," *American Political Science Review*, Vol. 90, No. 3 (September 1996), pp. 605–618.

12. Ibid.

generalizable variables that historians and researchers may have overlooked.

One way to avoid the risk of relying on a single historical analysis would be to follow the practice of Richard Smoke, who at the outset of his research, asked several historians to help him identify the best available accounts of each of the cases he planned to study. Later, Smoke obtained reviews of the first drafts of his cases from eight historians and made appropriate changes.[13]

Second, social scientists making use of even the best available historical studies of a case should not assume that they will provide answers to the questions they are asking. As emphasized in Chapter 3 on "The Method of Structured, Focused Comparison," the political scientist's research objectives determine the general questions to be asked of each case. The historian's research objectives and the questions addressed in his or her study may not adequately reflect those of subsequent researchers.[14] We may recall that historians have often stated that if history is approached from a utilitarian perspective, then it has to be rewritten for each generation. History does not speak for itself to all successive generations. When new problems and interests are brought to a study of history by later generations, the meaning and significance of earlier historical events to the present may have to be studied anew and reevaluated. Hence, the study of relevant historical experience very much depends on the specific questions one asks of historical cases.

One of the key tasks during the "soaking and poking" process is to identify the gaps in existing historical accounts. These gaps may include archival or interview evidence that has not been examined or that had previously been unavailable. They may also include the measurement of variables the researcher identified in phase one that historians have not measured or have not measured as systematically as the explanatory goals of subsequent researchers require. It is also possible that researchers can make use of technologies, such as computer-assisted content analysis, that were not available to scholars writing earlier historical accounts.

Third, having identified possible gaps in existing accounts, the re-

---

13. See the preface to Richard Smoke, *War: Controlling Escalation* (Cambridge, Mass.: Harvard University Press, 1977).

14. The different ways historians and political scientists tend to define the task of explanation and the different questions they often ask of available data is discussed in helpful detail in Deborah Larson, "Sources and Methods in Cold War History: The Need for a Theory-Based Archival Approach," in Colin Elman and Miriam Fendius Elman, eds., *Bridges and Boundaries: Historians, Political Scientists, and the Study of International Relations* (Cambridge, Mass.: MIT Press, 2001), pp. 327–350. The dangers of using studies by historians that may reflect their selection bias are noted also by Lustick, "History, Historiography, and Political Science."

searcher must reckon with the possibility that good answers to his or her questions about each case can be obtained only by going to original sources—archival materials, memoirs, oral histories, newspapers, and new interviews. In fact, political scientists studying international politics are increasingly undertaking this task. In doing so, however, they face the challenging task of weighing the evidentiary value of such primary sources.

Fourth, the researcher should not assume that going to primary sources and declassified government documents alone will be sufficient to find the answers to his or her research questions. The task of assessing the significance and evidentiary worth of such sources often requires a careful examination of contemporary public sources, such as daily media accounts of the developments of a case unfolding over time. Contemporary public accounts are certainly not a substitute for analysis of archival sources, but they often are an important part of contextual developments to which policymakers are sensitive, to which they are responding, or which they are attempting to influence. Classified accounts of the process of policymaking cannot be properly evaluated by scholars unless the public context in which policymakers operate is taken into account.[15] We have at times found students who have become intimately familiar with hard-to-get primary source materials of a case but who have only a vague sense of the wider context because they have not taken the relatively easy (but often time-consuming) step of reading the newspapers or journals from the period.[16]

---

15. The importance of studying contemporary journalistic sources in order to understand part of the context in which policymakers were operating became a central methodological procedure in Deborah Larson's research. In conjunction with thorough research into archival sources, Larson spent a great deal of time going through contemporary journalists' accounts of developments, a procedure which helped her to appreciate the impact of events that came to the attention of policymakers on their perceptions and responses. Careful study of the public context of private deliberations was useful in evaluating the evidentiary significance of archival sources. See Deborah Welch Larson, *The Origins of Containment* (Princeton, N.J.: Princeton University Press, 1985). Larson amplifies and illustrates different ways in which contemporary newspaper accounts help the investigator to discern important elements of the context in which policymakers operate. See Larson, "Sources and Methods in Cold War History."

16. One example comes from the work of one of the present authors, Andrew Bennett. In an unpublished study of the 1929 stock market crash for the Federal Reserve Board, he found by reading the newspapers of the period that there are strong reasons to question the often-cited argument that the crash was caused by excessive speculation on margin credit rates "as low as 10 percent," or the supposedly common practice of buying stocks by putting up only 10 percent of their value as equity. In fact, while no systematic data exists for the margins typically set, most newspaper accounts suggest that margins of 40 to 50 percent or higher were the norm. Banks offered to lower margin rates to 10 percent as an extraordinary step to try halt the crash, based

Finally, research on recent and contemporary U.S. foreign policy must be sensitive to the likelihood that important data may not be available and cannot be easily retrieved for research purposes, e.g., important discussions among policymakers that take place over the telephone or within internal e-mail and fax facilities—the results of which are not easily acquired by researchers.

The Risk of Over-Intellectualizing the Policy Process

When academic scholars attempt to reconstruct how and why important decisions were made, they tend to assume an orderly and more rational policymaking process than is justified. For example, overly complex and precise formal models may posit decision-making heuristics that are "too clever by half," or that no individual would actually utilize. Also, scholars sometimes succumb to the common cognitive bias toward univariate explanations—explanations in which there appears to be a single clear and dominating reason for the decision in question. Instead, analysts should be sensitive to the possibility that several considerations motivated the decision.

In fact, presidents and top-level executives often seek multiple payoffs from any decision they take. Leaders known for their sophistication and skill, such as Lyndon B. Johnson, use this strategy to optimize political gains from a particular decision. Disagreements among scholars as to the particular reason for why a certain action was taken often fail to take this factor into account.

Several considerations can enter into a decision in other ways as well. Particularly in a pluralistic political system in which a number of actors participate in policymaking, agreement on what should be done can emerge for different reasons. It is sufficient that members of the policymaking group agree only on *what* to do without having to agree on *why* to do it. In some situations, in fact, there may be a tacit agreement among members of the group that not all those who support the decision have to share the same reason or a single reason for doing so. To obtain sufficient

---

on the assumption that the crash was caused by a liquidity crisis as plunging stock values led to margin calls on stocks and forced sales of those stocks. The fact that this measure failed to stem the crash, and that bond purchases were strong during the crash, suggest that perhaps the crash was caused not so much by loose margin credit as by the classic bursting of a speculative bubble, and a revaluation of the relative value of stocks versus bonds. This explanation is more in line with modern theories of stock market behavior. In any event, a simple reading of the newspapers reveals that explanations of the crash cannot unproblematically accept that margins were typically 10 percent.

consensus on a decision may be difficult for various reasons, and sufficient time and resources may not be available for achieving a completely shared judgment in support of the decision. In any action-oriented group, particularly one that operates under time pressure, it is often enough to agree on what needs to be done. It may not be feasible or wise to debate until everyone agrees not merely on what decision to take but also the precise reasons for doing so.

Assessing the Evidentiary Value of Archival Materials

Scholars doing historical case studies must find ways of assessing the evidentiary value of archival materials that were generated during the policymaking process under examination. Similarly, case analysts making use of historical studies produced by other scholars cannot automatically assume that these investigators properly weighed the evidentiary significance of documents and interviews.

Scholars are not immune from the general tendency to attach particular significance to an item that supports their pre-existing or favored interpretation and, conversely, to downplay the significance of an item that challenges it. As cognitive dissonance theory reminds us, most people operate with a double standard in weighing evidence. They more readily accept new information that is consistent with an existing mind-set and employ a much higher threshold for giving serious consideration to discrepant information that challenges existing policies or preferences.

All good historians, it has been said, are revisionist historians. That is, historians must be prepared to revise existing interpretations when new evidence and compelling new interpretations emerge. Even seemingly definitive explanations are subject to revision. But new information about a case must be properly evaluated, and this task is jeopardized when a scholar is overly impressed with and overinterprets the significance of a new item—e.g., a recently declassified document—that emerges on a controversial or highly politicized subject.

Analytical or political bias on the scholar's part can lead to distorted interpretation of archival materials. But questionable interpretations can also arise when the analyst fails to grasp the context of specific archival materials. The importance of context in making such interpretations deserves more detailed analysis than can be provided here, so a few observations will have to suffice.

It is useful to regard archival documents as a type of purposeful communication. A useful framework exists for assessing the meaning and evidentiary worth of *what* is communicated in a document, speech, or interview. In interpreting the meaning and significance of what is said, the

analyst should consider *who* is speaking *to whom, for what purpose* and *under what circumstances.*[17] The evidentiary worth of what is contained in a document often cannot be reliably determined without addressing these questions. As this framework emphasizes, it is useful to ask what purpose(s) the document was designed to serve. How did it fit into the policymaking process? What was its relation to the stream of other communications and activities—past, present, and future?

It is also important to note the circumstances surrounding the document's release to the public, and to be sensitive to the possibility that documents will be selectively released to fit the political and personal goals of those officials who control their release. Much of the internal documentation on Soviet decision-making on the invasion and occupation of Afghanistan beginning in 1979, for example, was released by the government of Russian President Boris Yeltsin in the mid-1990s to embarrass the Soviet Communist Party, which was then on trial for its role in the 1991 Soviet coup attempt. Needless to say, the Yeltsin government did not release any comparable documents on its own ill-fated intervention in Chechnya in the mid-1990s.

In studying the outputs of a complex policymaking system, the investigator is well advised to work with a sophisticated model or set of assumptions regarding ways in which different policies are made in that system. For example, which actors and agencies are the most influential in a particular issue area? To whom does the leader turn for critical information and advice on a given type of policy problem? How do status differences and power variables affect the behavior of different advisers and participants in high-level policymaking?

Thus, it is advisable to observe a number of cautions in following the "paper trail" leading to a policy decision. Has a country's leader tipped his or her hand—at least in the judgment of participants in the pro-

---

17. This framework was initially developed and employed in a study that examined methods for inferring the intentions, beliefs, and other characteristics of a political elite from its propaganda by means of qualitative content analysis. See Alexander L. George, *Propaganda Analysis: A Study of Inferences Made from Nazi Propaganda in World War II* (Evanston, Ill.: Row, Peterson, 1959; and Westport, Conn.: Greenwood Press, 1973), pp. 107–121.

In a personal communication (March 26, 2000), Jeremi Suri drew on his own research experience to emphasize the need to distinguish between various types of archival materials. Personal correspondence and diaries of historical actors can be very helpful in developing understanding of their general beliefs about political life, particularly since such materials are often not designed to persuade others; such sources can reflect the emotions experienced at different junctures. Also, the "incoming files" of various reading matter insofar as it can be established that it was read, may throw light on the actor's ideology or cultural beliefs and the role they may play in policymaking.

cess—regarding what he or she will eventually decide? What effect does such a perception—or misperception—have on the views expressed or written by advisers? Are some of the influential policymakers bargaining with each other behind the leader's back regarding what advice and options to recommend in the hope and expectation that they can resolve their differences and protect their own interests?[18] What role did policymakers play in writing their own public speeches and reports, and to what extent do specific rhetorical formulations represent these top officials' own words rather than those of speech writers and other advisers?

It is well known that those who produce classified policy papers and accounts of decisions often wish to leave behind a self-serving historical record. One scholar who recently spent a year stationed in an office dealing with national security affairs witnessed occasions on which the written, classified record of important decisions taken was deliberately distorted for this and other reasons.[19] Diplomatic historian Stephen Pelz reminds us that "many international leaders take pains to disguise their reasoning and purposes, and therefore much of the best work on such figures as Franklin D. Roosevelt consists of reconstructing their assumptions, goals, and images of the world from a variety of sources."[20]

In assessing the significance of "evidence" that a leader has engaged in "consultation" with advisers, one needs to keep in mind that he or she may do so for several different reasons.[21] We tend to assume that he or she consults in order to obtain information and advice before making a final decision—i.e., to satisfy his or her "cognitive needs." But he or she may consult for any one or several other reasons. The leader may want to obtain emotional support for a difficult, stressful decision; or the leader may wish to give important advisers the feeling they have had an opportunity to contribute to the decision-making process so that they will be more likely to support whatever decision the president makes—i.e., to build consensus; or the leader may need to satisfy the expectation (generated by the nature of the political system and its political culture and

---

18. Some of these possibilities are among the various "malfunctions" of the policymaking system discussed in Alexander L. George, *Presidential Decisionmaking and Foreign Policy: The Effective Use of Information and Advice* (Boulder, Colo.: Westview Press, 1980), chap. 6.

19. This observation was provided by a scholar who must remain anonymous.

20. Stephen Pelz, "Toward A New Diplomatic History: Two and a Half Cheers for International Relations Methods," in Elman and Elman, eds., *Bridges and Boundaries*, p. 100.

21. This paragraph and the next one draw on George, *Presidential Decisionmaking*, pp. 81ff.

norms) that important decisions will not be made without the participation of all key actors who have some relevant knowledge, expertise, or responsibility with regard to the matter being decided; that is, the president hopes to achieve "legitimacy" for a decision by giving evidence that assures Congress and the public that it was well-considered and properly made. (Of course, a leader's consultation in any particular instance may combine several of these purposes.)

This last purpose—consultation—is of particular interest in the United States. The public wants to be assured that an orderly, rational process was followed in making important decisions. Consider the development in recent decades of "instant histories" of many important decisions by leading journalists on the basis of their interviews with policymakers shortly after the event. Knowing that the interested public demands to know how an important decision was made, top-level policymakers are motivated to conduct the decision process in ways that will enable them to assure the public later that the decision was made after careful multisided deliberation. Information to this effect is given to journalists soon after the decision is made. Since "instant histories" may be slanted to portray a careful, multidimensioned process of policymaking, the case analyst must consider to what extent such an impression is justified and how it bears on the evidentiary worth of the information conveyed in the instant history and in subsequent "insider" accounts of how and why a particular decision was made.

To weigh archival type material effectively, scholars need to be aware of these complexities. An excellent example of a study that captures the dynamics of decision-making is Larry Berman's interpretation of President Johnson's decision in July 1965 to put large-scale ground combat troops into Vietnam. Some archival sources suggest that Johnson employed a careful, conscientious version of "multiple advocacy" in which he thoughtfully solicited all views. But according to Berman's analysis, Johnson had already decided what he had to do and went through the motions of consultation for purposes of consensus-building and legitimization of his decision.[22]

In another example, many scholars assumed that President Dwight D. Eisenhower's policymaking system was highly formalistic and bureaucratic, a perception shared by important congressional and other critics at the time. Working with this image of Eisenhower's decision-making style, scholars could easily misinterpret the significance of archival sources generated by the *formal* track of his policymaking. Easily

22. Larry Berman, *Planning a Tragedy: The Americanization of the War in Vietnam* (New York: Norton, 1982).

overlooked was the *informal* track, which preceded and accompanied the formal procedures, awareness of which led Fred Greenstein to write about the "hidden hand style" by which Eisenhower operated.[23] Now, a more sophisticated way of studying Eisenhower's policymaking has developed that pays attention to both the formal and informal policy tracks and to the interaction between them.

The relevance and usefulness of working with an analytical framework that considers both tracks is, of course, not confined to studying the Eisenhower presidency. The workings of the informal track are not likely to become the subject of a written archival document. It is important to use interviews, memoirs, the media, etc., to obtain this valuable material.

Another aspect of the importance of a contextual framework for assessing the evidentiary worth of archival sources has to do with the hierarchical nature of the policymaking system in most governments. We find useful the analogy of a pyramid of several layers. Each layer, beginning with the bottom one, sends communications upwards (as well as sideways), analyzing available data on a problem and offering interpretations of its significance for policy. As one moves up the pyramid, the number of actors and participants grows smaller but their importance (potential, if not actual) increases. As one reaches the layer next to the top—the top being the president—one encounters a handful of key officials and top advisers. At the same time, we find that researchers at times interview officials who are too high in the hierarchy to have had close involvement in or detailed recall of the events under study. Often, lower-level officials who worked on an issue every day have stronger recollections of how it was decided than the top officials who actually made the decision but who focused on the issues in question only intermittently. However, a researcher must take into account that even well-informed lower-level officials often do not have a complete or fully reliable picture of how and why a decision was made—i.e., the "Rashomon" problem, when different participants in the process have different views as to what took place.

This layered pyramid produces an enormous number of communications and documents that the scholar must assess. The possibility of erroneous interpretation of the significance of archival material is enormous. How do sophisticated historians and other scholars cope with this problem? What cautions are necessary when examining archival sources on top-level policymaking? How does a researcher deal with the fact that much of the material coming to the top-level group of policymakers from

---

23. Fred I. Greenstein, *The Hidden-Hand Presidency: Eisenhower As Leader* (New York: Basic Books, 1982).

below is inconsequential? How does one decide which material coming from below to the top-level officials made a difference in the decision? How can one tell why he or she really decided as he or she did as against the justifications given for his or her decisions?

The analyst's search for documentary evidence on reasons behind top-level decisions can also run into the problem that the paper trail may end before final decisions are made. Among the reasons for the absence of reliable documentary sources on such decisions is the role that secrecy can play. Dean Rusk, Secretary of State during the Kennedy administration, later stated that secrecy "made it very difficult for many to reconstruct the Bay of Pigs operation, particularly its planning, because very little was put on paper. [Allen] Dulles, [Richard] Bissell, and others proposing the operation briefed us orally."[24]

No doubt there are important examples of scholarly disputes that illustrate these problems and indicate how individual analysts handled them. What general lessons can be drawn that would help train students and analysts? We have not yet found any book or major article that provides an adequate discussion of the problems of weighing the evidentiary worth of archival materials.[25] The most we can do, therefore, is to warn writers of historical case studies about some of these problems and to call attention to some of the methods historians and political scientists have employed in dealing with archival materials. Deborah Larson, for example, suggests that "to judge the influence of a memo written by a

---

24. As told to Richard Rusk in Daniel S. Papp, ed., *As I Saw It* (New York: Norton, 1990), cited by Richard Ned Lebow, "Social Science and History: Ranchers versus Farmers," in Elman and Elman, eds., *Bridges and Boundaries*, p. 132.

25. The most useful account we have found is the article by John D. Mulligan, "The Treatment of A Historical Source," *History and Theory*, Vol. 18, No. 2 (May 1979), pp. 177–196. Mulligan identifies various criteria historians employ for evaluating the authenticity, meaning, and significance of historical sources. He cites the observations on these issues made by a large number of distinguished historians and illustrates how each criterion applies to his own research, which focused on the importance of a correct evaluation of a primary source which sharply challenges accepted historical research on an aspect of the Civil War. This source was a personal letter, not a governmental document. Nonetheless, Mulligan's article illustrates the relevance of the framework we suggest, namely asking, "who says what to whom for what purpose in what circumstances?"

Also useful is the recent article by Cameron G. Thies, "A Pragmatic Guide to Qualitative Historical Analysis in the Study of International Relations," *International Studies Perspective*, Vol. 3, No. 4 (November 2002), pp. 351–372. This article includes a comprehensive list of sources that contributed to his essay. Readers may also want to consult the website "History Matters" <www.historymatters.gmu.edu.> which is designed for high school and college teachers of history. This website includes sections on "making sense of evidence" and "secrets of great history teachers."

lower-level official, one can look to see who initialed it. Of course, that a secretary of state initialed a memo does not prove that he read it, but it is a first step in analysis. Sometimes higher officials will make marginal comments—these can be quite important. Finally, paragraphs from memos written by lower officials sometimes appear in National Security Council policy memoranda."[26]

Problems in Evaluating Case Studies

Case writers should become familiar with the variety of critiques their work may face. The importance of understanding the history and context of a case makes the difficulties of critiquing qualitative research different from those of assessing quantitative work. Readers cannot easily judge the validity of the explanation of a case unless they possess a degree of independent knowledge of that case. This requires that reader-critics themselves possess some familiarity with the complexity of the case and the range of data available for studying it; knowledge of the existence of different interpretations offered by other scholars and of the status of the generalizations and theories employed by the case writer; and an ability to evaluate the case writer's use of counterfactual analysis or to provide plausible counterfactual analysis of their own. These are tough requirements for readers who must evaluate case studies, and simply to state these desiderata suffices to indicate that they are not easily met. Our own commentaries of case study research designs in the Appendix, "Studies That Illustrate Research Design," should be read with the caveat that we are not theoretical or historical experts on all the subjects of these studies. This is a problem also for those who review these books in academic journals.

Let us discuss some of the problems likely to be encountered by readers who attempt to evaluate case studies. Much of the preceding discussion is relevant to the task of evaluating case studies, and a few additional observations can be made.

The task of evaluating case studies differs depending on the research objective of the case. When the investigator's research objective is to explain a case outcome, the reader-critic must consider whether the case analyst has "imposed" a favored theory as the explanation. Have alternative theories that might provide an explanation been overlooked or inadequately considered? When the case writer pursues the different research objective of attempting to use case findings to "test" an existing

26. Letter from Deborah Welch Larson to Alexander L. George, April 10, 1999.

theory, there are several questions the reader-critic has to consider in deciding whether such a claim is justified. Does the case (or cases) constitute an easy or tough test of the theory? Do case findings really support the theory in question? Do they perhaps also support other theories the investigator has overlooked or inadequately considered?

Reader-critics must consider the possibility that the case-writer has overlooked or unduly minimized potentially important causal variables, or has not considered the possibility or likelihood that the phenomenon is subject to multiple conjunctural causation or is affected by equifinality.

These and other problems in using case studies to develop or test theories are also discussed in Chapter 6. They are referred to here in order to emphasize that case writers should be familiar with the variety of criticisms that can be and often are made of their work.

In addition, we urge that case writers accept the obligation to assist readers in evaluating whether their case analyses have met relevant methodological standards. To meet this requirement, case writers should go as far as reasonably possible to make the analyses they offer transparent enough to enable readers to evaluate them. Transparency of case studies must be closely linked with standards for case studies. These standards include (but are not limited to) providing enough detail to satisfy as much as possible the criteria of replicability and of the validity and reliability of the way in which variables are scored. Certainly these standards are often difficult to meet in case study research, but case writers can often do more to at least approximate them. We strongly concur with the admonition of Gary King, Robert Keohane, and Sidney Verba that *"the most important rule for all data collection is to report how the data were created and how we came to possess them."*[27]

In sum, case analysts should strive to develop and make use of appropriate rules for qualitative analysis. As argued in earlier chapters, however, the development of such guidelines should not be regarded as a matter of simply extending to qualitative analysis all of the standard conventions for quantitative analysis. Some of these conventions apply also to qualitative analysis, but guidelines for case studies must take into account the special characteristics of qualitative methodology.[28]

---

27. Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton, N.J.: Princeton University Press, 1994), p. 51. Emphasis in original.

28. For a detailed analysis of this position, see Gerardo L. Munck, "Canons of Research Design in Qualitative Analysis," *Studies in Comparative International Development*, Vol. 33, No. 3 (Fall 1998). The author provides a systematic and balanced assess-

Conclusion

The present book was in process of publication when we became aware of a new guidebook on how to make use of primary historical sources. The author, Marc Trachtenberg, has produced a superb manuscript which is in draft form for the time being. Its title is *Historical Method in the Study of International Relations.*

Himself a leading diplomatic historian, Trachtenberg joined the political science department at UCLA several years ago. He has succeeded in bringing together historical and political science approaches to the study of international relations. This book will be an invaluable source for students and professors who want to integrate the perspectives of history and political science for insightful research on foreign policies.

We will not attempt to summarize the rich materials he presents. The titles of several chapters may be noted: Chapter 3, "The Critical Analysis of Historical Texts"; and Chapter 5, "Working with Documents." A chapter is also provided on "Diplomatic History and International Relations Theory"; another chapter provides a detailed analysis of America's road to war in 1941.

Trachtenberg's treatment of these issues is unusually user-friendly. It is written in an engaging style. It will become standard text for research on foreign policy. Trachtenberg provides many incisive examples to illustrate his points.

We may also recall the statement that Trachtenberg made some time ago: "The basic methodological advice one can give is quite simple: documents are not necessarily to be taken at face value, and one has to see things in context to understand what they mean. One has to get into the habit of asking why a particular document was written—that is, what purpose it was meant to serve."[29]

We have stressed in the preceding pages the necessity to regard archival sources as being instances of purposive communication. This advice is strongly reinforced by Deborah Larson on the basis of her experience in conducting in-depth research in archival sources in preparing her book *Origins of Containment.*[30] A recent article by Larson helps to fill the gap regarding the proper use of archival sources, at least for research on U.S.

---

ment of the canons for qualitative research imbedded in King, Keohane, and Verba, *Designing Social Inquiry.*

29. In a letter to Alexander L. George (January 29, 1998), Marc Trachtenberg indicated that he is currently studying methods for assessing archival and other sources in research on international politics.

30. Larson, *The Origins of Containment.*

foreign policy. In it she emphasizes that it is important to understand the purpose of a document and the events leading up to it in order to correctly interpret its meaning. . . . The author of a memorandum or speaker at a meeting may be trying to ingratiate himself with superiors, create a favorable impression of himself, put himself on the record in case of leaks, or persuade others to adopt his preferred policy. Whatever his goals, we cannot directly infer the communicator's state of mind from his arguments without considering his immediate aims.[31]

Larson also notes that study of contemporary accounts in leading newspapers sometimes can be essential for ascertaining the context of documents. "News accounts can help to establish the atmosphere of the times, the purpose of speeches or statements, or the public reaction to a statement. Newspapers help to show what information policymakers had and provide clues as to what events they regarded as important. . . . In this way, newspapers help us to recapture the perspective of officials at the time."[32]

---

31. Deborah Welch Larson, "Sources and Methods in Cold War History," pp. 327–350.

32. Ibid. See also the project "Oral History Roundtables: The National Security Project," established in 1998 by Ivo H. Daalder and I.M. Destler, sponsored by the Brookings Institution and the Center for International and Security Studies at the University of Maryland. This series of roundtables, published periodically, brings together former officials specializing in foreign and security affairs to discuss specific historical problems in which they were involved. Daalder and Destler plan a final summary report.

Chapter 6

Phase Three: Drawing the Implications of Case Findings for Theory

Case study findings can have implications both for theory development and theory testing. On the inductive side of theory development, plausibility probes and studies of deviant cases can uncover new or omitted variables, hypotheses, causal paths, causal mechanisms, types, or interactions effects. Theory testing aims to strengthen or reduce support for a theory, narrow or extend the scope conditions of a theory, or determine which of two or more theories best explains a case, type, or general phenomenon. While many works on research methods and the philosophy of science emphasize theory testing more than theory development, we see both enterprises as essential to constructing good theories.

Case study findings can have implications for theory development and testing on three levels. First, they may establish, strengthen, or weaken historical explanations of a case. This is where within-case methods like process-tracing come into play. If a theory posits particular causal mechanisms as an explanation of a particular case, but these prove to be demonstrably absent, then the theory is greatly weakened as an explanation for this case, though there is still the possibility of measurement error or omitted variables.

Yet a modified historical explanation of a case may not add to explanations of other cases that are dissimilar in some respects. Establishing the general applicability of a new or modified explanation of a case requires showing that it accurately explains other cases. Conversely, invalidating an existing theory as an explanation of one case does not necessarily imply that the theory poorly explains other, dissimilar cases; indeed, the existing theory may have earlier demonstrated a strong ability to ex-

plain cases.[1] Whereas some earlier approaches assumed or demanded that a new theory subsume or explain all of the phenomena explained by its predecessors, we do not require that this always be so. A new theory may be superior in explaining only some of the cases explained by its predecessor, or even only one case, while being inapplicable to others.

Second, and more generally, the finding that a theory does or does not explain a case may be generalized to the type or class of cases (e.g., deterrence) of which this case is a member. Here, the generalization depends on the precision and completeness with which the class of cases has been defined and the degree to which the case exemplifies the class. Generalization to cases not studied always entails some risk of mistaken inferences because they may differ from the case or cases studied in the values of potentially causal variables omitted from the theoretical framework.

Third and most broadly, case study findings may in some circumstances be generalized to neighboring cells in a typology, to the role of a particular variable in dissimilar cases, or even to all cases of a phenomenon. Here overgeneralization is a risk, since the analyst is generalizing cases that differ in the value of variables that have been already identified as causally related to the outcome. This is why case study researchers usually limit themselves to narrow and well-specified contingent generalizations about a type.[2] Still, some cases may constitute particularly strong tests of theories, allowing generalization beyond the particular cases studied.

This chapter looks at each of these kinds of generalization, first in theory development and then in theory testing. It concludes that improved historical explanations of individual cases are the foundation for drawing wider implications from case studies, as they are a necessary condition for any generalizations beyond the case. Contingent or typological generalizations are often the most useful kind of theoretical

---

1. The Bayesian approach to theory choice is one means of weighting the confidence we should place in an existing theory versus a new competing theory. Briefly, in the Bayesian approach, we increase our prior estimate of the likely truth of a theory when we encounter evidence that is likely only if the theory is true and unlikely if alternative explanations are true. This relies, however, on subjective prior probabilities that researchers assign to the truth of competing theories. The Bayesian defense of this practice is that as evidence accumulates, differences in the prior probabilities that different researchers assign to theories will "wash out" as new evidence forces researchers' confidence in theories to converge. For arguments on both sides of this issue, see John Earman, *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* (Cambridge, Mass.: MIT Press, 1992).

2. David Collier and James Mahoney, "Insights and Pitfalls: Selection Bias in Qualitative Research," *World Politics*, Vol. 49, No. 1 (October 1996), pp. 59–91.

conclusions from case studies, as they build on and go beyond improved historical explanations but present limited risks of extending these conclusions to causally dissimilar cases. Findings that can be extended to different types of cases are less common, and often must be stated as only loose generalizations. However, they can be important turning points in research programs, drawing attention toward avenues for future research.

Theory Development

The development of theory via case studies should be distinguished from the deductive development of theory. Deductive methods can usefully develop entirely new theories or fill the gaps in existing theories; case studies can test deductive theories and suggest new variables that need to be incorporated. (The literature on deterrence, as noted below, provides an excellent example of this process.) But theory development via case studies is primarily an inductive process. This section highlights the usefulness of deviant cases for inductively identifying new variables or causal mechanisms. (Plausibility probes, which we do not discuss here, also focus directly on the goal of theory development, by aiming at clearer specification of a theory and its variables and by attempting to better identify which cases might prove most valuable for theory building.)

THEORY DEVELOPMENT AND HISTORICAL EXPLANATION OF SINGLE CASES

The outcome in a deviant case may prove to have been caused by variables that had been previously overlooked but whose effects are well known from other research. This leads to an improved historical explanation of the case, but not necessarily to any new generalizations from the case, unless the case is one in which the previously overlooked variables were not expected to have any effect.

An inductively derived explanation of a case can also involve more novel theories and variables. In this context, researchers are frequently advised not to develop a theory from evidence and then test it against the same evidence; facts cannot test or contradict a theory that is constructed around them. In addition, using the same evidence to create and test a theory also exacerbates risks of confirmation bias, a cognitive bias toward affirming one's own theories that has been well documented both in laboratory experiments and in the practices of social scientists.[3]

However, it is valid to develop a theory from a case and then test the

3. For a study that indicates that social scientists' explanations for the failures of their predictions appear to be biased in favor of their initial theories, see Philip Tetlock,

theory against additional evidence from the case that was not used to derive the theory. This makes the theory falsifiable as an explanation for the case, and can circumvent confirmation biases. Researchers, even when they are fairly expert on a case and its outcome (or the value of its dependent variable), are often ignorant of the detailed processes through which the outcome arose.[4] As a researcher begins to delve into primary sources, there are many opportunities to reformulate initial explanations of a case in ways that accommodate new evidence and also predict what the researchers should find in evidence they have not yet explored or had not even thought to look for. Researchers can also predict what evidence they should find in archives before these are made accessible or in interviews before they are carried out.[5] Indeed, in testing a historical explanation of a case, the most convincing procedure is often to develop an explanation from data in the case and then test it against other evidence in the case; otherwise, the only recourse is to test the explanation in other cases that differ in ways that may prevent generalization back to the original case.

THEORY DEVELOPMENT AND CONTINGENT GENERALIZATIONS

The study of a deviant case can lead a researcher to identify a new type of case. As we discuss in Chapter 11, this process can take place through a "building block" approach, with new case studies identifying subtypes or the causal processes that apply to a subtype of cases. Each case study thus contributes to the cumulative refinement of contingent generalizations on the conditions under which particular causal paths occur, and fills out the cells or types of a more comprehensive theory.

Historians often view efforts to generalize from historical case studies with suspicion. Yet one can generalize from unique cases by treating

---

"Theory-Driven Reasoning about Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions?" *American Journal of Political Science*, Vol. 43, No. 2 (April 1999), pp. 348–349. Researchers should also be on guard against other cognitive biases, including the bias toward over-confidence in one's causal theories, a preference for uni-causal explanations, and a tendency toward assuming that causes resemble consequences in terms of scale, scope, or complexity.

4. Researchers also often find that their preliminary knowledge of the values of the independent and dependent variables is mistaken, particularly if it is based on news accounts or secondary sources that do not use precise definitions. Thus even these variables can provide some use-novelty for researchers; however, as we note in our chapter on congruence testing, tests of the congruence of independent and dependent variables, even with the advantage of use-novelty, are challenging and often less conclusive than process-tracing tests.

5. William Wohlforth suggests this practice in "Reality Check: Revising Theories of International Politics in Response to the End of the Cold War," *World Politics*, Vol. 50, No. 4 (July 1998), pp. 650–680.

them as members of a class or type of phenomenon; that is, as instances of alliance formation, deterrence, war initiation, negotiation, peacekeeping, war termination, revolution, and so on. This is often followed by distinguishing subclasses of each of the phenomena. Researchers can also develop "concatenated" theories by dividing a complex causal process into its specific component theories, or sequential stages, focusing on particular policy instruments or the views of designated actors. For example, Alexander George and Richard Smoke divided deterrence theory into a number of more specific theories which deterrence comprises: commitment theory, initiation theory, and response theory.[6] Similarly, Bruce Jentleson, Ariel Levite, and Larry Berman broke down protracted military interventions into sequential stages and the differing dynamics of getting in, staying in, and getting out.[7] Such designations help identify subtypes of undertakings and phenomena that occur repeatedly throughout history which can be grouped together and studied as a class or subclass of similar events. This can be done through statistical analysis when a sufficiently large number of cases of a particular phenomenon is available, or through qualitative analysis of a small number of instances.

Where should one draw the line in developing ever more finely grained types and subtypes? As Sidney Verba put it many years ago:

> To be comparative, we are told, we must look for generalizations or covering laws that apply to all cases of a particular type. But where are the general laws? Generalizations fade when we look at particular cases. We add intervening variable after intervening variable. Since the cases are few in number, we end up with an explanation tailored to each case. The result begins to sound quite idiographic or configurative... In a sense we have come full circle. . . . As we bring more and more variables back into our analysis in order to arrive at any generalizations that hold up across a series of political systems, we bring back so much that we have a "unique" case in its configurative whole.[8]

Yet Verba did not conclude that the quest for theory and generalization is infeasible. Rather, the solution to this apparent impasse is to formulate the idiosyncratic aspects of the explanation for each case in terms

---

6. Alexander L. George and Richard Smoke, *Deterrence in American Foreign Policy: Theory and Practice* (New York: Columbia University Press, 1974). See also the discussion of George and Smoke in the Appendix, "Studies That Illustrate Research Design."

7. Bruce Jentleson, Ariel Levite, and Larry Berman, eds., *Foreign Military Intervention: The Dynamics of Protracted Conflict* (New York: Columbia University Press, 1992). This book is discussed in the Appendix, "Studies That Illustrate Research Design."

8. Sidney Verba, "Some Dilemmas of Political Research," *World Politics*, Vol. 20, No. 1 (October 1967), pp. 113–114.

of general variables. "The 'unique historical event' cannot be ignored," Verba notes, "but it must be considered as one of a class of events even if it happened only once."[9]

One criterion that helps determine where to draw the line in the proliferation of subtypes is the notion of "leverage"—the desirability of having theories that explain as many dependent variables as possible with as few simple independent variables as possible. This is not the same as parsimony, or simplicity of theories. We agree with Verba and his co-authors Gary King and Robert Keohane that parsimony is "an assumption . . . about the nature of the world: it is assumed to be simple . . . but we believe [parsimony] is only occasionally appropriate . . . theory should be just as complicated as all our evidence suggest."[10]

The recognition that even unique cases can contribute to theory development strengthens the linkage between history and political science. Some of the particular qualities of each case are inevitably lost in the process of moving from a specific to a more general explanation. The critical question, however, is whether the loss of information and simplification jeopardizes the validity and utility of the theory. This question cannot be answered abstractly or *a priori.* Much depends upon the sensibility and judgment of the investigator in choosing and conceptualizing variables and also in deciding how best to describe the variance in each of the variables. The latter task in particular—the way in which variations for each variable are formulated—may be critical for capturing the essential features of "uniqueness." For this reason, investigators should develop the categories for describing the variance in each of their variables inductively, via detailed examination of how the value of a particular variable differs across many different cases.

THEORY DEVELOPMENT AND GENERALIZING ACROSS TYPES

The most general kind of finding from a deviant case is the specification of a new concept, variable, or theory regarding a causal mechanism that affects more than one type of case and possibly even all instances of a phenomenon. This specification of new concepts or variables, as Max Weber noted, is often one of the most important contributions of research.[11] Charles Darwin's theory of evolution, for example, was sparked by a small number of cases (particularly the small differences between

---

9. Ibid.

10. Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton, N.J.: Princeton University Press, 1994), pp. 29, 20.

11. Marianne Weber, *Max Weber: A Biography,* trans. Harry Zohn (New Brunswick,

finches on the South American mainland and those on the Galapagos Islands), but it posited new causal mechanisms of wide relevance to biological and even social systems.

When a deviant case leads to the specification of a new theory, the researcher may be able to generalize about how the newly identified mechanism may play out in different contexts, or he or she may only be able to suggest that it should be widely relevant. As an example of the former, Andrew Bennett, Joseph Lepgold, and Danny Unger undertook a study of burden sharing in the 1991 Gulf War partly because several countries' sizeable contributions to the Desert Storm coalition contradicted the collective action theories that then dominated the literature on alliances and would have predicted more free-riding. The authors found that pressure from the United States, the coalition leader, explained the large contributions by allies dependent on the United States for their security, most notably Germany and Japan. While pressure from a powerful state is not a novel hypothesis in explaining international behavior, the finding suggested that the collective action hypothesis was generally less determinative in alliance behavior than had been argued. While the temptation of free riding grows as one state becomes more powerful relative to others, so does the ability of the powerful state to coerce dependent allies as well. As these forces offset one another, other factors—domestic politics and institutions, the nature of the public good of alliance security, and so on—help tilt the balance toward or away from a contribution. In short, the authors developed fairly detailed contingent generalizations on how the understudied factor of alliance dependence would play out in different contexts.[12]

Theory Testing

When theories are fairly well developed, researchers can use case studies for theory testing. The goal here is rarely to refute a theory decisively, but rather to identify whether and how the scope conditions of competing theories should be expanded or narrowed. This is a challenging process: when a theory fails to fit the evidence in a case, it is not obvious whether the theory fails to explain the particular case, fails to explain a whole class of cases, or does not explain any cases at all. Should we blame a theory's

---

N.J.: Transaction Press, 1988), p. 278, cited in David Laitin, "Disciplining Political Science," *American Political Science Review*, Vol. 89, No. 2 (June 1995), p. 455.

12. See Chapter 11 for a more detailed discussion of this research as an example of typological theory. See also Andrew Bennett, Joseph Lepgold, and Danny Unger, "Burden-Sharing in the Persian Gulf War," *International Organization*, Vol. 48, No. 1 (Winter 1994), pp. 39–75.

failure on a flaw in the theory's internal logic or on contextual conditions that rendered the theory inapplicable (which would require only a narrowing of the theory's scope conditions to exclude the anomalous case), or on some combination of the two? We should not be too quick to reject general theories on the basis of one or a few anomalous cases, as these theories may still explain other cases very well. Conversely, there is a danger of too readily retaining a false theory by narrowing its scope conditions to exclude anomalous cases, or by adding additional variables to the theory to account for anomalies.

An additional difficulty in theory testing is that tests are partly dependent on the causal assumptions of theories themselves. For example, theories that posit simple causal relations, such as necessity, sufficiency, or linearity can be falsified by a single case (barring measurement error). Theories are harder to test if they posit more complex causal relations, such as equifinality and interactions effects. Still, such theories, which are often the kind that most interest case study researchers, may be subjected to strong tests if they assume high-probability (but not necessarily deterministic) relations between variables and posit a manageably small number of variables, interactions, and causal paths. Theories are hardest to subject to empirical tests if they involve the most complex types of causal relations, or what might be called "enigmatic" causality: complex interactions among numerous variables, low-probability relations between variables, and endogeneity problems or feedback effects. Such theories are difficult to test even with large numbers of cases to study. Although a single case can disprove a deterministic assertion, even many cases cannot *falsify* a probabilistic claim—it is only increasingly unlikely to be true if it fails to fit a growing number of cases.

While theories need to be developed into a testable form, a theory should not be forced into predictions beyond its scope; this leads to the creation of an easily discounted "straw man" version of the theory. A test could also be too tough if countervailing variables mask the causal effects of the variable under study.[13] Of course, researchers frequently disagree on whether a theory is being forced to "stick its neck out" sufficiently far, or whether it is being pushed into predictions beyond its rightful scope.[14] If an empirical test is beyond the domain of phenomena to which the the-

13. See Stephen W. Van Evera, *Guide to Methods for Students of Political Science* (Ithaca, N.Y.: Cornell University Press, 1997), p. 34.

14. See, for example, Colin Elman, "Horses for Courses: Why Not Neorealist Theories of Foreign Policy," *Security Studies*, Vol. 6, No. 1 (Autumn 1996), pp. 7–53. Elman critiques neorealist theories for claiming to eschew any testable predictions on individual states' foreign policies.

ory has been applied, then findings inconsistent with the theory limit its scope rather than falsify it.

How can a researcher avoid too readily rejecting or narrowing the scope conditions of a theory that is in fact accurate, or accepting or broadening the scope conditions of a theory that is in fact false or inapplicable? There are no infallible criteria for addressing all of the complications of generalizing the results of a case study's theory tests. A key consideration, however, is the issue of how tough an empirical test a case poses for a theory: How strongly do the variables predict the case's outcome, and how unique are the predictions the theory makes for the case?[15]

TESTING COMPETING EXPLANATIONS OF CASES

An explanation of a case is more convincing if it is more unique, or if the outcome it predicts "could not have been expected from the best rival theory available."[16] If a phenomenon has not previously received wide study, a theory can only make a rather weak claim to being the "best" explanation. For closely studied phenomena, however, the finding that a case fits only one explanatory theory is powerful evidence that the theory best explains the case. Of the five hypotheses considered in the study of burden-sharing in the 1991 Gulf War noted above (balance of threat, alliance dependence, collective action, domestic politics, and policymaking institutions) *only* the alliance dependence hypothesis fit the outcome and process of the German and Japanese contributions to the coalition. This highlighted the power of alliance dependence, since the variables identified by all the other hypotheses militated against this outcome.

---

15. In a similar formulation, Stephen Van Evera suggests that the probity of an empirical test depends on the certainty and uniqueness of the predictions a theory makes regarding the test. "Hoop tests" are those in which the predictions of a theory are certain but not unique. Failing such a test is damaging to a theory, but passing it is not definitive. "Smoking gun tests" are those in which a theory is unique but not certain. Passing such a test is strong corroboration, but failing it does not undermine a theory. "Doubly decisive" tests, when predictions are both unique and certain, are those in which either passage or failure is definitive. (Van Evera gives the example here of a bank camera, which can both convict those guilty of robbery and exculpate the innocent.) "Straw-in-the-wind" tests, with predictions of low certainty and uniqueness, are not definitive regardless of the outcome. See Van Evera, *Guide to Methods*, pp. 31–32.

16. Colin Elman and Miriam Fendius Elman, "How Not to be Lakatos Intolerant: Appraising Progress in IR Research," *International Studies Quarterly*, Vol. 46, No. 2 (June 2002), p. 240, citing M. Carrier, "On Novel Facts: A Discussion of Criteria for Non-Ad-Hocness in the Methodology of Scientific Research Programs," *Zeitschrift für allgemeine Wissenschaftstheorie*, Vol. 19, No. 2 (1988), pp. 205–231. In the philosophy of science, a theory that makes a unique prediction is said to have achieved "background theory novelty."

In testing competing historical explanations of a case, then, it is important to find instances where explanations make unique predictions about the process or outcome of the case. An excellent example of this is Scott Sagan's work on the safety of nuclear weapons from accidental or unauthorized use.[17] Sagan treats the safety of nuclear weapons as a subclass of the ability of complex organizations to manage hazardous technology. The latter problem has been addressed in two major theories: Charles Perrow's normal accidents theory, and the high reliability theory developed by a group of Berkeley scholars.[18] Neither of these two organizational theories had addressed the specific problem of nuclear weapons safety, but Sagan argues they each have implications for this issue.

Sagan notes that both theories often make ambiguous predictions.[19] Neither theory excludes the possibility of a serious accident, though the normal accident theory is more pessimistic. There is considerable overlap between the two in their predictions on the nuclear weapons cases of interest to Sagan, but he finds the theories to be at odds in several important respects. Sagan notes that "many of the specific conditions that the high reliability theorists argue will promote safety will actually reduce safety according to the normal accidents theorists." Conversely, he argues, the safety requirements posited by the high reliability school are impossible to implement in the view of normal accidents theorists.[20]

Sagan identifies historical situations, including several aspects of the Cuban Missile Crisis, in which the theories make different predictions about the level of safety achieved and the means through which it was attained.[21]

Sagan notes that his goal was to "deduce what each theory should predict about specific efforts to prevent the ultimate safety system failure—an accidental nuclear war—and then compare these predictions to the historical experiences of U.S. nuclear weapons command and control.

---

17. Scott D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton N.J.: Princeton, University Press, 1993).

18. Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (New York: Basic Books, 1984); and Todd LaPorte and Paula Consolini, "Working in Practice but Not in Theory: Theoretical Challenges of 'High Reliability Organizations,'" *Journal of Public Administration Research and Theory*, Vol. 1, No. 1 (January 1991), pp. 19–47.

19. Sagan, *The Limits of Safety*, pp. 13, 49.

20. Ibid., p. 45.

21. Ibid., p. 51. A debate on Sagan's book was later published between Todd LaPorte, a leading adherent of the "high reliability" school and Charles Perrow, the founder of the "normal accidents" school. A comment on their exchange is provided by Scott Sagan in *Journal of Contingencies and Crisis Management*, Vol. 2, No. 4 (December 1994), pp. 205–240.

Which theory provides better predictions of what happened and more compelling explanations of why it happened? Which theory leads to the discovery of more novel facts and new insights? Which one is therefore a better guide to understanding?"[22] Sagan concludes that on the whole, the normal accidents school provides more accurate answers to these questions in the case of the Cuban Missile Crisis.

Sagan's reasoning is as follows: given that there have been no accidental nuclear wars, one can focus on the performance of the two theories in predicting and explaining the serious—though not catastrophic—failures in the safety of nuclear weapons that have occurred. An interesting feature is Sagan's effort to construct a tough test for the normal accidents theory in the impressive U.S. safety record with nuclear weapons, which appears to conform more closely to the optimistic predictions of high reliability theorists. That U.S. leaders attach high priority to avoiding accidental nuclear war, U.S. nuclear forces personnel are isolated from society and subject to strict military discipline, and the United States has adequate resources to spend on the safety of its nuclear weapons also favors the validity of the high reliability theory and poses a tough test for the normal accidents theory. Sagan nonetheless concludes on the basis of detailed process-tracing evidence that the lesser safety failures and near misses that did occur are comprehensible only in terms of the warnings of the normal accidents school. By arriving at this finding even in a very tough test, Sagan creates a convincing basis for generalizing beyond his cases to U.S. nuclear weapons safety as a whole.

TESTING CONTINGENT GENERALIZATIONS

To test contingent or typological generalizations, scholars must clearly specify the scope or domain of their generalizations. To what range of institutional settings, cultural contexts, time periods, geographic settings, and situational contexts do the findings apply? Here again, typological theorizing, as discussed in Chapter 11, provides a ready means for specifying the configurations of variables or the types to which generalizations apply. Tests of contingent generalizations can then consist of examining cases within the specified domain of the theory to see if their processes and outcomes are as the theory predicts. Conversely, researchers can test for cases beyond the specified scope conditions of the theory to determine if these scope conditions might be justifiably broadened.

The proper boundaries of contingent generalizations are a frequent subject of contention among theorists. An illuminating example concerns Theda Skocpol's study of social revolutions in France, Russia, and

22. Sagan, *The Limits of Safety*, p. 49.

China.[23] Barbara Geddes critiques Skocpol's analysis by arguing that in several Latin American countries, the causes of revolution that Skocpol identified were present, but no revolutions occurred, while in other countries in the region, revolutions took place even in the absence of the preconditions Skocpol noted.[24] Skocpol was careful to make her theory contingent, however, clearly indicating in her introduction and conclusion that her theory is not a general theory of revolutions, but a theory of revolutions in wealthy agrarian states that had not experienced colonial domination. Skocpol in fact explicitly states that her argument does not apply to three cases that Geddes raises (Mexico in 1910, Bolivia in 1952, and Cuba in 1959), so these cases do not contravene the scope conditions that Skocpol outlines.[25] A more appropriate critique of Skocpol would point out cases that fit within the domain Skocpol defined but that do not fit her theory, or criticize directly the way in which Skocpol defined the domain of her theory.[26]

GENERALIZING ACROSS TYPES: TOUGH TESTS AND MOST-LIKELY, LEAST-LIKELY, AND CRUCIAL CASES

It is difficult to judge the probative value of a particular test relative to the weight of prior evidence behind an existing theory. Harry Eckstein argues that "crucial cases" provide the most definitive type of evidence on a theory. He defines a crucial case as one "that *must closely fit* a theory if one is to have confidence in the theory's validity, or conversely, *must not fit* equally well with any rule contrary to that proposed." He adds that "in a crucial case it must be extremely difficult, or clearly petulant, to dismiss any finding contrary to the theory as simply 'deviant' (due to chance, or the operation of unconsidered factors)."[27]

---

23. Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia, and China* (Cambridge, U.K.: Cambridge University Press, 1979).

24. Barbara Geddes, "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics," in James A. Stimson, ed., *Political Analysis*, Vol. 2 (Ann Arbor: University of Michigan Press, 1990). This example of the Skocpol-Geddes debate is from Collier and Mahoney, "Insights and Pitfalls," pp. 80–82.

25. Collier and Mahoney, "Insights and Pitfalls," p. 81.

26. Along these lines, as Chapter 2 notes, there is a debate over whether new democracies should be excluded from tests of democratic peace theories. Some view the exclusion of new democracies from statistical tests of these theories as an arbitrary way to rescue the theories from anomalous findings. Others view the exclusion as legitimate, arguing that the causal mechanisms that create a democratic peace are only very weakly established in states in transition to democracy.

27. Harry Eckstein, "Case Studies in Political Science," in Fred Greenstein and Nelson Polsby, eds., *Handbook of Political Science*, Vol. 7 (Reading, Mass.: Addison-Wesley, 1975), p. 118. McKeown suggests that in this regard case study researchers use an in-

Eckstein notes the difficulties in identifying such crucial cases when theories and their predictive consequences are not precisely stated, but notes that the foremost problem is that truly crucial cases rarely occur in nature or the social world. Therefore, he suggests the alternative of tough tests which entail studying most-likely and least-likely cases. In a most-likely case, the independent variables posited by a theory are at values that strongly posit an outcome or posit an extreme outcome. In a least-likely case, the independent variables in a theory are at values that only weakly predict an outcome or predict a low-magnitude outcome. Most-likely cases, he notes, are tailored to cast strong doubt on theories if the theories do not fit, while least-likely cases can strengthen support for theories that fit even cases where they should be weak.

Many case study researchers have identified the cases they choose for study as most-likely or least-likely cases, but it is necessary to be explicit and systematic in determining this status. One must consider not only whether a case is most or least likely for a given theory, but whether it is also most or least likely for alternative theories. One useful means of doing so, as noted in Chapter 11 on typological theory, is to include a typological table that shows the values of variables in the case or cases studied for competing hypotheses. Such a table helps the researcher and reader identify which variables in a case may favor alternative theories, and helps the researcher to address systematically whether alternative theories make the same or different predictions on processes and outcomes in a given case.

In general, the strongest possible supporting evidence for a theory is a case that is least likely for that theory but most likely for all alternative theories, and one where the alternative theories collectively predict an outcome very different from that of the least-likely theory. If the least-likely theory turns out to be accurate, it deserves full credit for a prediction that cannot also be ascribed to other theories (though it could still be spurious and subject to an as-yet undiscovered theory). This might be called a toughest test case.[28] Theories that survive such a difficult test may prove to be generally applicable to many types of cases,

---

formal version of Bayesian logic. Timothy J. McKeown, "Case Studies and the Statistical World View," *International Organization*, Vol. 53, No. 1 (Winter 1999), pp. 161–190.

28. Similarly, Margaret Mooney Marini and Burton Singer define the "gross strength" of a causal inference on the role of a variable X as the overall evidence consistent with "X causes Y," and they define the "net strength" on X as the gross strength of X discounted by the gross strength of alternative variables and their underlying theories. Margaret Mooney Marini and Burton Singer, "Causality in the Social Sciences," in Clifford Clogg, ed., *Sociological Methodology*, Vol. 18 (1998), pp. 347–409. See also James Caporoso, "Research Design, Falsification, and the Qualitative-Quantitative Divide," *American Political Science Review*, Vol. 89, No. 2 (June 1995), p. 458.

as they have already proven their robustness in the presence of countervailing mechanisms.

The best possible evidence for weakening a theory is when a case is most likely for that theory and for alternative theories, and all these theories make the same prediction. If the prediction proves wrong, the failure of the theory cannot be attributed to the countervailing influence of variables from other theories (again, left-out variables can still weaken the strength of this inference). This might be called an easiest test case. If a theory and all the alternatives fail in such a case, it should be considered a deviant case and it might prove fruitful to look for an undiscovered causal path or variable. A theory's failure in an easiest test case calls into question its applicability to many types of cases.

One example of a theory that failed an easy test case comes from Arend Lijphart's study of the Netherlands, which cast doubt on David Truman's theory of "cross-cutting cleavages."[29] Truman had argued that mutually reinforcing social cleavages, such as coterminous class and religious cleavages, would lead to contentious politics, while cross-cutting cleavages would lead to cooperative social relations. In the Netherlands, however, Lijphart found a case with essentially no cross-cutting cleavages but a stable and cooperative democratic political culture. This cast doubt on Truman's theory not just for the Netherlands, but more generally.

Cases usually fall somewhere in between being most and least likely for particular theories, and so pose tests of an intermediate degree of difficulty. Short of finding toughest or easiest test cases, researchers should be careful to specify, for each alternative hypothesis, where the case at hand lies on the spectrum from most to least likely for that theory, and when the theory predicts outcomes that complement or contradict other theories' predictions.

For example, Graham Allison's study of the Cuban Missile Crisis, *Essence of Decision*, is in some respects a strong test case for the rational actor model, a moderate test of the organizational process model, and a strong test of the bureaucratic politics model.[30] However, it is not the strongest

---

29. This example comes from Ronald Rogowski, "The Role of Theory and Anomaly in Social-Scientific Inference," *American Political Science Review*, Vol. 89, No. 2 (June 1995), pp. 467–468; the referenced works are Arend Lijphart, *The Politics of Accommodation: Pluralism and Democracy in the Netherlands* (Berkeley: University of California Press, 1975); and David Truman, *The Governmental Process: Political Interest and Public Opinion* (New York: Knopf, 1951).

30. Graham Allison and Philip Zelikow, *Essence of Decision: Explaining the Cuban Missile Crisis*, 2nd ed. (Longman, N.Y.: Longman, 1999).

possible test of any model and just how strong it is depends on which of Allison's research questions is under consideration.

Let us consider the first two of Allison's three research questions as examples. On the question of "Why did the Soviet Union place missiles in Cuba?" rational actor considerations should have been strong given the clear strategic stakes. Organizational processes should not have been very strong because the Soviet Union was taking the initiative and had time to adapt its procedures. Bureaucratic politics should have been of moderate importance given the stakes involved for Soviet military budgets and missions. On the question of "Why did Kennedy react as he did?" rational actor considerations were constrained by the incomplete information and short time period, but strengthened by the president's direct involvement. On the other hand, the nature of the crisis favored U.S. decision-making that approximates the rational actor model. Organizational processes were a moderate constraint—the president's personal involvement could and did modify procedures, but the short time available limited possible adaptations. Bureaucratic politics should have been constrained by the president's role and the overriding importance of national concerns (rather than parochial institutional concerns). One could add details on what makes each question a most- or least- likely case for each of the models, but the general point is that many contextual factors must be taken into account and that they rarely all point in the same direction on the high likelihood of one theory and the low likelihood of others.

It is important to note that a case in which one variable is at an extreme value is not necessarily a definitive test. Rather, if the variables of competing explanations make the same prediction and are not at extreme values, this may represent an easy test that provides only weak evidence for the importance of the extreme variable. Such easy tests are not very probative, and if they are incorrectly used to infer strong support for a theory, they may constitute a problem of selection bias. Such a case may be more useful for the heuristic purpose of identifying the outsized causal mechanisms related to the extreme variable.

Conclusion

Generalizing the results of case studies is not a simple function of the number or diversity of cases studied. A researcher may study diverse cases that prove to have no common patterns, so that only unique historical explanations of each case are possible. Alternatively, a researcher may study a few cases or even one case and uncover a new causal mechanism that proves applicable to a wide range of cases. Single cases can also cast

doubt on theories across a wide range of scope conditions, as Arend Lijphart's study of the Netherlands demonstrates. These extremes of a complete inability to generalize from a case and a warrant for broad generalizations from a single case are relatively infrequent. More common is the opportunity to use case study findings to incrementally refine middle-range contingent generalizations, either by broadening or narrowing their scope or introducing new types and subtypes through the inclusion of additional variables. Such refinements draw on both within-case analyses, which help test historical explanations of cases, and cross-case comparisons, which help identify the domains to which these explanations extend. This interplay among within-case analyses and comparative methods is the hallmark of typological theorizing, a subject to which we return in Chapter 11.