CHAPTER 1

Hypotheses, Laws, and Theories: A User's Guide

What Is a Theory?

Definitions of the term "theory" offered by philosophers of social science are cryptic and diverse.[1] I recommend the following as a simple framework that captures their main meaning while also spelling out elements they often omit.

Theories are general statements that describe and explain the

1. Most posit that theories explain phenomena and leave it at that. The elements of an explanation are not detailed. See, for example, Brian Fay and J. Donald Moon, "What Would an Adequate Philosophy of Social Science Look Like?" in Michael Martin and Lee C. McIntyre, eds., *Readings in the Philosophy of Social Science* (Cambridge: MIT Press, 1994), p. 26: a social theory is a "systematic, unified explanation of a diverse range of social phenomena." Likewise Earl Babbie, *The Practice of Social Research*, 7th ed. (Belmont, Calif.: Wadsworth, 1995), p. 40: "A theory is a systematic explanation for the observations that relate to a particular aspect of life." See also Kenneth Waltz, quoted in note 9. Each leaves the components of an explanation unspecified.

Leaving even explanation unmentioned is W. Phillips Shively, *The Craft of Political Research*, 3d ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1990): "A theory takes a set of similar things that happen—say, the development of party systems in democracies—and finds a common pattern among them that allows us to treat each of these different occurrences as a repeated example of the same thing" (p. 2).

causes or effects of classes of phenomena. They are composed of causal laws or hypotheses, explanations, and antecedent conditions. Explanations are also composed of causal laws or hypotheses, which are in turn composed of dependent and independent variables. Fourteen definitions bear mention:

law

An observed regular relationship between two phenomena. Laws can be deterministic or probabilistic. The former frame invariant relationships ("if *A* then always *B*"). The latter frame probabilistic relationships ("if *A* then sometimes *B*, with probability *X*"). Hard science has many deterministic laws. Nearly all social science laws are probabilistic.

Laws can be causal ("*A* causes *B*") or noncausal ("*A* and *B* are caused by *C*; hence *A* and *B* are correlated but neither causes the other").[2] Our prime search is for causal laws. We explore the possibility that laws are noncausal mainly to rule it out, so we can rule in the possibility that observed laws are causal.[3]

2. Generic laws (which might be causal or noncausal) should be stated in associative language ("if *A*, then *B*," or "the greater *A*, the greater *B*", or "the higher *A*, the smaller *B*", etc.). Causal laws can also be framed with causal language ("*A* causes *B*").

3. Causal laws can assume four basic causal patterns: direct causation ("*A* causes *B*"), reverse causation ("*B* causes *A*"), reciprocal causation ("*A* causes *B* and *B* causes *A*"), and self-undermined causation ("*A* causes *B* and *B* lessens *A*"). Hypotheses, discussed below, can assume the same formats. To establish a specific causal relationship ("*A* causes *B*"), we must rule out the possibility that an observed relationship between *A* and *B* is spurious ("*C* causes *A* and *B*") or reverse-causal ("*B* causes *A*"). We may also investigate whether reciprocal causation or self-undermined causation is at work.

hypothesis	A conjectured relationship between two phenomena.[4] Like laws, hypotheses can be of two types: causal ("I surmise that *A* causes *B*") and noncausal ("I surmise that *A* and *B* are caused by *C*; hence *A* and *B* are correlated but neither causes the other").
theory	A causal law ("I have established that *A* causes *B*") or a causal hypothesis ("I surmise that *A* causes *B*"), together with an explanation of the causal law or hypothesis that explicates how *A* causes *B*. Note: the term "general theory" is often used for more wide-ranging theories, but all theories are by definition general to some degree.
explanation	The causal laws or hypotheses that connect the cause to the phenomenon being caused, showing how causation occurs. ("*A* causes *B* because *A* causes *q*, which causes *r*, which causes *B*.")
antecedent condition[5]	A phenomenon whose presence activates or

4. This follows P. McC. Miller and M. J. Wilson, *A Dictionary of Social Science Methods* (New York: John Wiley, 1983), p. 58: "[A hypothesis is] a conjecture about the relationships between two or more concepts." Carl Hemple uses "hypothesis" more broadly, to include conjectures about facts as well as relationships. Thus, for Hempel, descriptive conjectures (for instance, estimates of the height of the Empire State Building or the size of the national debt) are also hypotheses. See Carl G. Hempel, *Philosophy of Natural Science* (Englewood Cliffs, N.J.: Prentice-Hall, 1966), p. 19. I use the term "propositions" to refer to what Hempel calls "hypotheses": thus, for me, propositions can be hypotheses or descriptive conjectures. Babbie, *Practice of Social Research*, also uses "hypothesis" broadly (see p. 49); under "hypothesis" he includes predictions inferred from hypotheses (which I call "predictions," "observable implications," or "test implications" of theory).

5. The term is from Carl G. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press, 1965), pp. 246–47 and passim. The term "antecedent" merely means that the condition's presence pre-

magnifies the action of a causal law or hypothesis. Without it causation operates more weakly ("*A* causes some *B* if *C* is absent, more *B* if *C* is present"—e.g., "Sunshine makes grass grow, but causes large growth only in fertilized soil") or not at all ("*A* causes *B* if *C* is present, otherwise not"—e.g., "Sunshine makes grass grow, but only if we also get some rainfall").

We can restate an antecedent condition as a causal law or hypothesis. ("*C* causes *B* if *A* is present, otherwise not"—e.g., "Rainfall makes grass grow, but only if we also get some sunshine").

Antecedent conditions are also called "interaction terms," "initial conditions," "enabling conditions," "catalytic conditions," "preconditions," "activating conditions," "magnifying conditions," "assumptions," "assumed conditions," or "auxiliary assumptions."

variable — A concept that can have various values, e.g., the "degree of democracy" in a country or the "share of the two-party vote" for a political party.

independent variable (IV) — A variable framing the causal phenomenon of a causal theory or hypothesis. In the hypothesis "literacy causes democracy," the degree of literacy is the independent variable.

---

cedes the causal process that it activates or magnifies. Antecedent conditions need not precede the arrival of the independent variable onto the scene; they can appear after the appearance of high values on the independent variable that they activate or magnify.

dependent variable (DV)	A variable framing the caused phenomenon of a causal theory or hypothesis. In the hypothesis "literacy causes democracy," the degree of democracy is the dependent variable.
intervening variable (IntV)	A variable framing intervening phenomenon included in a causal theory's explanation. Intervening phenomena are caused by the IV and cause the DV.[6] In the theory "Sunshine causes photosynthesis, causing grass to grow," photosynthesis is the intervening variable.
condition variable (CV)[7]	A variable framing an antecedent condition. The values of condition variables govern the size of the impact that IVs or IntVs have on DVs and other IntVs. In the hypothesis "Sunshine makes grass grow, but only if we also get some rainfall," the amount of rainfall is a condition variable.
study variable (SV)	A variable whose causes or effects we seek to discover with our research. A project's study variable can be an IV, DV, IntV, or CV.
prime hypothesis	The overarching hypothesis that frames the relationship between a theory's independent and dependent variables.

6. Whether a specific variable is dependent, independent, or intervening depends on its context and changes with context, as with *A* in these statements: (1) "*A* causes *B*": *A* is the independent variable; (2) "*Q* causes *A*": *A* becomes the dependent variable; and (3) "*Q* causes *A*, and *A* causes *B*": *A* becomes an intervening variable.

7. Condition variables are also known as "suppressor" variables, meaning that controlling values on these variables suppresses irregular variance between independent and dependent variables. See Miller and Wilson, *Dictionary of Social Science Methods*, p. 110.

explanatory hypothesis	The intermediate hypotheses that constitute a theory's explanation.[8]
test hypothesis	The hypothesis we seek to test. Also called the "research hypothesis."

Note: a theory, then, is nothing more than a set of connected causal laws or hypotheses.[9]

We can always "arrow-diagram" theories, like this:

$$A \rightarrow q \rightarrow r \rightarrow B$$

In this diagram A is the theory's independent variable, B is the dependent variable. The letters q and r indicate intervening vari-

8. These last four terms—"condition variable," "study variable," "prime hypothesis," and "explanatory hypothesis"—are my own nominations to fill word-gaps in the lexicon.

9. For a different view see Kenneth N. Waltz, *Theory of International Politics* (Reading, Mass.: Addison-Wesley, 1979), pp. 2, 5. To Waltz, theories are not "mere collections of laws" but rather the "statements that explain them" (p. 5). These statements include "theoretical notions," which can take the form of concepts or assumptions. I prefer my definition to Waltz's because all explanations for social science laws that I find satisfying can be reduced to laws or hypotheses. His definition of "explanation" also lacks precision because it leaves the prime elements of an explanation unspecified.

For a third meaning, more restrictive than mine, see Christopher H. Achen and Duncan Sindal, "Rational Deterrence Theory and Comparative Case Studies," *World Politics* 41 (January 1989): 147: A theory is "a very general set of propositions from which others, including 'laws,' are derived." Their definition omits modestly general ideas that I call theories.

Nearer my usage is Carl Hempel: "Theories . . . are bodies of systematically related hypotheses." Carl G. Hempel, "The Function of General Laws in History," in Martin and McIntyre, *Readings in the Philosophy of Social Science*, p. 49. Likewise Miller and Wilson, *Dictionary of Social Science Methods:* "[A theory is] a set of integrated hypotheses designed to explain particular classes of events" (p. 112). Similar are Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press, 1994), p. 99: "Causal theories are designed to show the causes of a phenomenon or set of phenomena" and include "an interrelated set of causal hypotheses. Each hypothesis specifies a posited relationship between variables."

ables and comprise the theory's explanation. The proposal "$A \rightarrow B$" is the theory's prime hypothesis, while the proposals that "$A \rightarrow q$," "$q \rightarrow r$," and "$r \rightarrow B$" are its explanatory hypotheses.

We can add condition variables, indicating them by using the multiplication symbol, "×."[10] Here C is a condition variable: the impact of A on q is magnified by a high value on C and reduced by a low value on C.

$$\begin{array}{l} A \rightarrow q \rightarrow r \rightarrow B \\ \times \\ C \end{array}$$

An example would be:

$$\begin{array}{lll} \text{Amount of} & \text{Amount of} & \text{Amount of} \\ \text{sunshine} \rightarrow & \text{photosynthesis} \rightarrow & \text{grass growth} \\ \times & & \\ \text{Amount of} & & \\ \text{rainfall} & & \end{array}$$

One can display a theory's explanation at any level of detail. Here I have elaborated the link between r and B to show explanatory variables s and t.

$$\begin{array}{l} A \rightarrow q \rightarrow r \rightarrow s \rightarrow t \rightarrow B \\ \times \\ C \end{array}$$

One can extend an explanation to define more remote causes. Here remote causes of A (Y and Z) are detailed:

10. The multiplication sign is used here only to indicate that the CV magnifies the impact of the IV, not to mean that the CV literally multiplies the impact of the IV (although it might).

$$Y \rightarrow Z \rightarrow A \rightarrow q \rightarrow r \rightarrow s \rightarrow t \rightarrow B$$
$$\times$$
$$C$$

We can detail the causes of condition variables, as here with the cause of *C*:

$$Y \rightarrow Z \rightarrow A \rightarrow q \rightarrow r \rightarrow s \rightarrow t \rightarrow B$$
$$\times$$
$$X \rightarrow C$$

There is no limit to the number of antecedent conditions we can frame. Here more conditions (*D*, *u*, *v*) are specified.

$$\begin{array}{ccccccccccc} Y & \rightarrow & Z & \rightarrow & A & \rightarrow & q & \rightarrow & r & \rightarrow & s & \rightarrow & t & \rightarrow & B \\ & & & & \times & & & & & & \times & & & & \\ & & & & X & \rightarrow & C & & & & u & & & & \\ & & & & \times & & & & & & \times & & & & \\ & & & & D & & & & & & v & & & & \end{array}$$

One can add more avenues of causation between causal and caused variables. Here two chains of causation between *A* and *B* (running through intervening variables *f* and *g*) are added, to produce a three-chain theory:

$$\begin{array}{ccccccccccccccc} & & & & & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & f & \rightarrow & \\ Y & \rightarrow & Z & \rightarrow & A & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & g & \rightarrow & B \\ & & & & & \rightarrow & q & \rightarrow & r & \rightarrow & s & \rightarrow & t & \rightarrow & \\ & & & & \times & & & & & & \times & & & & \\ & & & X & \rightarrow & C & & & & & u & & & & \\ & & & & \times & & & & & & \times & & & & \\ & & & & D & & & & & & v & & & & \end{array}$$

A "theory" that cannot be arrow-diagrammed *is not a theory* and

needs reframing to become a theory. (According to this criteria much political science "theory" and "theoretical" writing is not theory.)

What Is a Specific Explanation?

Explanations of specific events (particular wars, revolutions, election outcomes, economic depressions, and so on) use theories and are framed like theories. A good explanation tells us what specific causes produced a specific phenomenon and identifies the general phenomenon of which this specific cause is an example. Three concepts bear mention:

specific explanation	An explanation cast in specific terms that accounts for a distinctive event. Like a theory, it describes and explains cause and effect, but these causes and effects are framed in singular terms. (Thus "expansionism causes aggression, causing war" is a theory; "German expansionism caused German aggression, causing World War II" is a specific explanation.) Specific explanations are also called "particular explanations" (as opposed to "general explanations.")

Specific explanations come in two types. The second type ("generalized specific explanation") is more useful:

nongeneralized specific explanation	A specific explanation that does not identify the theory that the operating cause is an example of. ("Germany caused World War II." The explanation does not answer the ques-

	tion "of what is Germany an example?")[11]
generalized specific explanation	A specific explanation that identifies the theories that govern its operation.[12] ("German expansionism caused World War II." The operating cause, "German expansionism," is an example of expansionism, which is the independent variable in the hypothesis "expansionism causes war.")

Specific explanations are composed of causal, caused, intervening, and antecedent phenomena:[13]

causal phenomenon (CP)	The phenomenon doing the causing.
caused phenomenon (OP)	The phenomenon being caused.
intervening phenomena (IP)	Phenomena that form the explanation's explanation. These are caused by the causal phenomenon and cause the outcome phenomenon.
antecedent phenomena (AP)	Phenomena whose presence activates or magnifies the causal action of the causal and/or explanatory phenomena.[14]

11. Such explanations rest on implicit theories, however, as Carl Hempel has explained. See Hempel, "Function of General Laws in History."

12. The theories thus identified are sometimes termed the "warrants" of the argument or explanation. See Wayne C. Booth, Gregory G. Colomb, and Joseph M. Williams, *The Craft of Research* (Chicago: University of Chicago Press, 1995), pp. 90–92, 111–31. "The warrant of an argument is its general principle, an assumption or premise that bridges the claim and its supporting evidence" (ibid., p. 90).

13. Specific explanations are composed of singular phenomena that represent specific values on variables, not of variables themselves. As such they are "phenomena," not "variables." On assessing specific explanations see "How Can Specific Events Be Explained," in this chapter.

14. These last seven terms—"specific explanation," "nongeneralized specific explanation," "generalized specific explanation," "causal phenomenon," "caused phenomenon," "intervening phenomenon," and "antecedent phenomenon"—

We arrow-diagram specific explanations the same way we do theories:

A theory	Expansionism → Aggression → War
A generalized specific explanation	German expansionism → German aggression → World War II
A nongeneralized specific explanation	Germany → Outbreak of fighting on September 1, 1939 → World War II

What Is a Good Theory?

Seven prime attributes govern a theory's quality.

1. A good theory has *large explanatory power.* The theory's independent variable has a large effect on a wide range of phenomena under a wide range of conditions. Three characteristics govern explanatory power:

Importance. Does variance in the value on the independent variable cause large or small variance in the value on the dependent variable?[15] An important theory points to a cause that has a large impact—one that causes large variance on the dependent vari-

---

are my suggested labels for these concepts. Others use "explanandum phenomenon" for the caused phenomenon, and "explanans" for a generalized explanation and its components (the causal, intervening, and antecedent phenomena). See, for example, Hempel, *Philosophy of Natural Science,* p. 50. (In Hempel's usage only generalized specific explanations comprise an explanans—nongeneralized specific explanation do not.)

15. A theory's importance can be measured in "theoretical" or "dispersion" terms. A theoretical measure of importance asks: how many units of change in the value on the dependent variable are caused by a unit of change in the value on the independent variable? (How many additional votes can a candidate gain by spending an additional campaign dollar on television ads?) A dispersion measure asks: what share of the DV's total variance in a specific data set is caused by variance of this IV? (What percentage of the variance in the votes received by various congressional candidates is explained by variance in their television spending?) I use "importance" in the former sense, to refer to theoretical importance. See Christopher H. Achen, *Interpreting and Using Regression* (Beverly Hills: Sage, 1982), pp. 68–77.

able. The greater the variance produced, the greater the theory's explanatory power.

Explanatory range. How many classes of phenomena does variance in the value on the theory's independent variable affect, hence explain? The wider the range of affected phenomena, the greater the theory's explanatory power. Most social science theories have narrow range, but a few gems explain many diverse domains.[16]

Applicability. How common is the theory's cause in the real world? How common are antecedent conditions that activate its operation? The more prevalent the causes and conditions of the theory, the greater its explanatory power.[17] The prevalence of

16. Karl Deutsch used the terms "combinatorial richness" and "organizing power" for attributes similar to what I call explanatory range, with "combinatorial richness" expressing "the range of combinations or patterns that can be generated from" a model, and "organizing power" defining the correspondence of the theory or model to phenomena other than those it was first used to explain. Karl Deutsch, *The Nerves of Government* (New York: Free Press, 1966), pp. 16–18. Examples of social science theories with wide explanatory range include Mancur Olson's theory of public goods, Robert Jervis's offense-defense theory of war and arms racing, Stanislav Andreski's military-participation ratio (MPR) explanation for social stratification, and Stephen Walt's balance-of-threat theory of alliances. See Mancur Olson, *The Logic of Collective Action* (Cambridge: Harvard University Press, 1971); Robert Jervis, "Cooperation under the Security Dilemma," *World Politics* 30 (January 1978): 167–214; Stanislav Andreski, *Military Organization and Society* (Berkeley: University of California Press, 1971), pp. 20–74; and Stephen M. Walt, *The Origins of Alliances* (Ithaca: Cornell University Press, 1987), pp. 17–33.

17. Even causes that produce powerful effects can have little explanatory power if these causes are rare in the real world, or if they require rare antecedent conditions to operate. Conversely, causes that produce weaker effects can have greater explanatory power if the cause and its antecedent conditions are common. Thus great white shark attacks are often lethal, but they explain few deaths because they are scarce in the real world. The cause is strong but rare, hence it explains little. Sunburn is less lethal but explains more death (through skin cancer) because it is more common. Likewise, scuba diving is often lethal if hungry great white sharks are around, but scuba diving explains few deaths because divers avoid shark-infested waters. The cause is powerful under the right conditions (hungry sharks nearby), but these conditions are rare, hence the cause explains few events. Sunburn explains more deaths because it does not require rare conditions to produce its harmful effects.

these causes and conditions in the past govern its power to explain history. Their current and future prevalence govern its power to explain present and future events.

2. Good theories elucidate by simplifying. Hence a good theory is *parsimonious*. It uses few variables simply arranged to explain its effects.

Gaining parsimony often requires some sacrifice of explanatory power, however. If that sacrifice is too large it becomes unworthwhile. We can tolerate some complexity if we need it to explain the world.

3. A good theory is "*satisfying*," that is, it satisfies our curiosity. A theory is unsatisfying if it leaves us wondering what causes the cause proposed by the theory. This happens when theories point to familiar causes whose causes, in turn, are a mystery. A politician once explained her election loss: "I didn't get enough votes!" This is true but unsatisfying. We still want to know why she didn't get enough votes.

The further removed a cause stands from its proposed effect, the more satisfying the theory. Thus "droughts cause famine" is less satisfying than "changes in ocean surface temperature cause shifts in atmospheric wind patterns, causing shifts in areas of heavy rainfall, causing droughts, causing famine."

4. A good theory is *clearly framed.* Otherwise we cannot infer predictions from it, test it, or apply it to concrete situations.

A clearly framed theory fashions its variables from concepts that the theorist has clearly defined.

A clearly framed theory includes a full outline of the theory's explanation. It does not leave us wondering how *A* causes *B*. Thus "changes in ocean temperature cause famine" is less complete than "changes in ocean temperature cause shifts in atmospheric wind patterns, causing shifts in areas of heavy rainfall, causing droughts, causing famine."

A clearly framed theory includes a statement of the antecedent conditions that enable its operation and govern its impact. Other-

wise we cannot tell what cases the theory governs and thus cannot infer useful policy prescriptions.

Foreign policy disasters often happen because policymakers apply valid theories to inappropriate circumstances. Consider the hypothesis that "appeasing other states makes them more aggressive, causing war." This was true with Germany during 1938–39, but the opposite is sometimes true: a firm stand makes the other more aggressive, causing war. To avoid policy backfires, therefore, policymakers must know the antecedent conditions that decide if a firm stand will make others more or less aggressive. Parallel problems arise in all policymaking domains and highlight the importance of framing antecedent conditions clearly.

5. A good theory is in principle *falsifiable.* Data that would falsify the theory can be defined (although it may not now be available).[18]

Theories that are not clearly framed may be nonfalsifiable because their vagueness prevents investigators from inferring predictions from them.

Theories that make omnipredictions that are fulfilled by all observed events are also nonfalsifiable. Empirical tests cannot corroborate or infirm such theories because all evidence is consistent with them. Religious theories of phenomena have this quality: happy outcomes are God's reward, disasters are God's punishment, cruelties are God's tests of our faith, and outcomes that elude these broad categories are God's mysteries. Some Marxist arguments share this omni-predictional trait.[19]

6. A good theory *explains important phenomena:* it answers ques-

18. Discussing this requirement of theory is Hempel, *Philosophy of Natural Science,* pp. 30–32.

19. For other examples see King, Keohane, and Verba, *Designing Social Inquiry,* p. 113, mentioning Talcott Parsons's theory of action and David Easton's systems' analysis of macropolitics. On Easton see also Harry Eckstein, "Case Study and Theory in Political Science," in Fred I. Greenstein and Nelson W. Polsby, eds., *Handbook of Political Science,* vol. 7, *Strategies of Inquiry* (Reading, Mass.: Addison-Wesley, 1975), p. 90.

tions that matter to the wider world, or it helps others answer such questions. Theories that answer unasked questions are less useful even if they answer these questions well. (Much social science theorizing has little real-world relevance and thus fails this test.)

7. A good theory has *prescriptive richness.* It yields useful policy recommendations.

A theory gains prescriptive richness by pointing to manipulable causes, since manipulable causes might be controlled by human action. Thus "capitalism causes imperialism, causing war" is less useful than "offensive military postures and doctrines cause war," even if both theories are equally valid, because the structure of national economies is less manipulable than national military postures and doctrines. "Teaching chauvinist history in school causes war" is even more useful, since the content of national education is more easily adjusted than national military policy.

A theory gains prescriptive richness by identifying dangers that could be averted or mitigated by timely countermeasures. Thus theories explaining the causes of hurricanes provide no way to prevent them, but they do help forecasters warn threatened communities to secure property and take shelter.

A theory gains prescriptive richness by identifying antecedent conditions required for its operation (see point 4). The better these conditions are specified the greater our ability to avoid misapplying the theory's prescriptions to situations that the theory does not govern.

How Can Theories Be Made?

There is no generally accepted recipe for making theories.[20] Some scholars use deduction, inferring explanations from more

20. Arguing the impossibility of a recipe is Hempel, *Philosophy of Natural Science,* pp. 10–18. Also see Milton Friedman, *Essays in Positive Economics* (Chicago: University of Chicago Press, 1953): constructing hypotheses "is a creative act of inspiration, intuition, invention . . . the process must be discussed in psychologi-

general, already-established causal laws. Thus much economic theory is deduced from the assumption that people seek to maximize their personal economic utility. Others make theories inductively: they look for relationships between phenomena; then they investigate to see if discovered relationships are causal; then they ask "of what more general causal law is this specific cause-effect process an example?" For example, after observing that clashing efforts to gain secure borders helped cause the Arab-Israeli wars, a theorist might suggest that competition for security causes war.[21]

Nine aids to theory-making bear mention. (The first eight are inductive, the last is deductive.)

1. We can examine "outlier" cases, that is, cases poorly explained by existing theories.[22] Unknown causes must explain their outcomes. We try to identify these causes by examining the case.

Specifically, to make a new theory we select cases where the phenomenon we seek to explain is abundant but its known causes are scarce or absent. Unknown causes must be at work. These causes will announce themselves as unusual characteristics of the case and as phenomena that are associated with the dependent variable within the case. We nominate these phenomena as candidate causes.[23] We also cull the views of people who experienced

---

cal, not logical, categories; studied in autobiographies and biographies, not treatises on scientific method; and promoted by maxim and example, not syllogism or theorem" (p. 43). On the subject of theory-making see also Shively, *Craft of Political Research*, pp. 163–66, where Shively notes the possibility of creating theories by induction, deduction, and borrowing theories from other fields.

21. From there the theorist could move further by returning to deduction, for instance, deducing that conditions that intensify competition for security—such as an advantage for the offensive on the battlefield—are also causes of war.

22. Such cases lie furthest from the regression line expressing the relationship between the dependent variable and its known causes; hence the term "outlier" cases. Another term for exploring outlier cases is "deviant-case analysis." See Arend Lijphart, "Comparative Politics and the Comparative Method," *American Political Science Review* 65 (September 1971): 692.

23. For example, India is a democracy with a low level of public literacy. Literacy

the case or know it well and nominate their explanations as candidate causes.

To infer a theory's antecedent conditions (CVs), we select cases where the dependent variable's causes are abundant but the dependent variable is scarce or absent. This suggests that unknown antecedent conditions are absent in the case. Study of the case may identify them.

2. The "method of difference" and "method of agreement" (proposed by John Stuart Mill)[24] can serve as aids to inductive theory-making. In the method of difference the analyst compares cases with similar background characteristics and different values on the study variable (that is, the variable whose causes or effects we seek to discover), looking for other differences between cases. We nominate these other cross-case differences as possible causes of the study variable (if we seek to discover its causes) or its possible effects (if we seek its effects). We pick similar cases to reduce the number of candidate causes or effects that emerge: the more similar the cases, the fewer the candidates, making real causes and effects easier to spot.[25] Likewise, in the method of

---

is an established cause of democracy, hence India is an "outlier" case, falling far from the regression line expressing the relationship between degree of democracy (the dependent variable) and levels of literacy (the independent variable). Exploring the India case will uncover causes of democracy that operate independently of literacy and in addition to it.

24. John Stuart Mill, *A System of Logic*, ed. J. M. Robson (Toronto: University of Toronto Press, 1973), chap. 8, "Of the Four Methods of Experimental Inquiry," pp. 388–406.

25. An example of using paired method-of-difference case studies for theory-making is Morris P. Fiorina, *Congress: Keystone of the Washington Establishment* (New Haven: Yale University Press, 1977), chap. 4, pp. 29–37. Fiorina sought to explain why marginal congressional districts ("swing" districts where Democrats and Republicans compete evenly in congressional elections) were disappearing. To generate hypotheses he compared two districts highly similar in character but different in result: one district had always been and remained marginal, the other had changed from marginal to nonmarginal during the 1960s. He nominated the key cross-district difference that he observed (greater constituent servicing by the congressional incumbent in the newly nonmarginal

agreement the analyst explores cases with different characteristics and similar values on the study variable, looking for other similarities between the cases, and nominating these similarities as possible causes or effects of the variable.[26]

3. We can select cases with extreme high or low values on the

---

district) as a possible cause of the general decline of marginality. The growth of government, he theorized, had created opportunities for incumbents to win the voters' favor by performing constituent service, and this bolstered incumbents who seized the opportunity.

I also had an early social science adventure inferring a hypothesis by method-of-difference case comparison (although I was oblivious of J. S. Mill at the time). In 1969 I sought to explain why black political mobilization remained low in the rural Deep South even after the passage of the 1965 Voting Rights Act. I inferred an explanation—holding that economic coercion by whites was retarding black mobilization—partly from Delphi-method interviews (see note 27) but also from a method-of-difference comparison.

I started by comparing two very similar black-majority Mississippi counties. Holmes and Humphries counties were virtual twins on nearly all socioeconomic dimensions except one: blacks had won county-wide elections in Holmes but lost badly in next-door Humphries. This spurred my search for a second difference between them. It was easy to spot. Holmes had the Mileston project, a community of black landowners who bought small farms through the New Deal Farm Security Administration in the 1940s. Humphries had nothing similar. As a result Holmes had far more black landowners than Humphries. Further investigation (process tracing) revealed that these landowners had played a key role in building Holmes County's black political organization. Interviews further suggested that fear of eviction among black tenant farmers deterred their political participation throughout Mississippi, and the Mileston farmers were emboldened to participate by their freedom from fear of eviction. A large-*n* test using all twenty-nine black-majority Mississippi counties then found a significant correlation between measures of black freedom from economic coercion and black political mobilization. This further corroborated the hypothesis that economic coercion depressed black political mobilization in the Mississippi black belt and suggested that such coercion might explain low levels of black mobilization across the rural Deep South.

The results of this study are summarized in Lester M. Salamon and Stephen Van Evera, "Fear, Apathy, and Discrimination: A Test of Three Explanations of Political Participation," *American Political Science Review* 67 (December, 1973): 1288–1306. (Unfortunately, our article omits my Holmes county interview and process-tracing data. Still wet behind the ears, I assumed that only large-*n* tests were valid and never thought to present Holmes county as a case study.)

26. The method of difference is more efficient when the characteristics of available cases are quite homogeneous (that is, when most aspects of most cases are

study variable (SV) and explore them for phenomena associated with it. If values on the study variable are very high (if the SV phenomenon is present in abundance), its causes and effects should also be present in unusual abundance, standing out against the case background. If values on the SV are very low (if the SV phenomenon is nearly absent), its causes and effects should also be conspicuous by their absence.

4. We can select cases with extreme within-case variance in the value on the study variable and explore them for phenomena that covary with it. If values on the study variable vary sharply, its causes and effects should also vary sharply, standing out against the more static case background.

5. Counterfactual analysis can aid inductive theorizing. The analyst examines history, trying to "predict" how events would have unfolded had a few elements of the story been changed, with a focus on varying conditions that seem important and/or manipulable. For instance, to explore the effects of military factors on the likelihood of war, one might ask: "How would pre-1914 diplomacy have evolved if the leaders of Europe had not believed that conquest was easy?" Or, to explore the importance of broad social and political factors in causing Nazi aggression: "How might the 1930s have unfolded had Hitler died in 1932?" The greater the impact of the posited changes, the more important the analysis.

When analysts discover counterfactual analyses they find persuasive, they have found theories they find persuasive, since all counterfactual predictions rest on theories. (Without theories the analyst could not predict how changed conditions would have changed events.) If others doubt the analysis (but cannot expose fatal flaws in it), all the better: the theory may be new, hence a real discovery. At this point the analyst has only to frame the theory in

---

similar). The method of agreement is preferred when the characteristics of cases are heterogeneous (that is, when most aspects of most cases are different).

a general manner so that predictions can be inferred from it and tested. The analyst should ask: "What general causal laws are the dynamics I assert examples of?" The answer is a theory.

Counterfactual analysis helps us recognize theories, not make them. Theories uncovered by counterfactual analysis must exist in the theorist's subconscious before the analysis; otherwise the theorist could not construct the counterfactual scenario. Most people believe in more theories than they know. The hard part is to bring these theories to the surface and express them in general terms. Counterfactual analysis aids this process.

6. Theories can often be inferred from policy debates. Proponents of given policies frame specific cause-effect statements ("If communism triumphs in Vietnam, it will triumph in Thailand, Malaysia, and elsewhere") that can be framed as general theories ("Communist victories are contagious: communist victory in one state raises the odds on communist victory in others"; or, more generally, "Revolution is contagious; revolution in one state raises the odds on revolution in others"). We can test these general theories. Such tests can in turn help resolve the policy debate. Theories inferred in this fashion are sure to have policy relevance, and they merit close attention for this reason.

7. The insights of actors or observers who experienced the event one seeks to explain can be mined for hypotheses. Those who experience a case often observe important unrecorded data that is unavailable to later investigators. Hence they can suggest hypotheses that we could not infer from direct observation alone.[27]

27. I used this technique—the "Delphi method"—to infer a hypothesis explaining why black political mobilization remained low in the rural Deep South even after the passage of the 1965 Voting Rights Act. At that time (1969) political scientists widely assumed that low black political mobilization stemmed from black political apathy. I thought the skill of local organizers might be key. Interviews, however, revealed that rural black community leaders doubted both theories. They instead argued that fear of white coercion deterred black participation, and freedom from coercion helped explain pockets of black political mobiliza-

8. Large-*n* data sets can be explored for correlations between variables. We nominate discovered correlations as possible cause-effect relationships. This method is seldom fruitful, however. A new large-*n* data set is usually hard to assemble, but if we rely on existing data sets, our purview is narrowed by the curiosities of previous researchers. We can only explore theories that use variables that others have already chosen to code.

9. We can fashion theories by importing existing theories from one domain and adapting them to explain phenomena in another.[28] Thus students of misperception in international relations and students of mass political behavior have both borrowed theories from psychology. Students of military affairs have borrowed theories from the study of organizations. Students of international systems have borrowed theories (e.g., oligopoly theory) from economics.

How Can Theories Be Tested?

We have two basic ways to test theories: experimentation and observation. Observational tests come in two varieties: large-*n* and case study. Thus, overall we have a universe of three basic testing methods: experimentation, observation using large-*n* analysis, and observation using case-study analysis.[29]

---

tion. Further investigation found substantial evidence to support their argument. (This hypothesis also emerged from a method-of-difference comparison of two Mississippi counties. See note 25.)

28. Suggesting this technique is Shively, *Craft of Political Research*, p. 165.

29. Deduction supplies a fourth way to evaluate theories. Using deduction to evaluate the hypothesis that *a* causes *b*, we would ask if *a* and *b* are examples of more general phenomena (*A* and *B*) that are already known to cause each other. If so, we can deduce that, since *A* causes *B*, and *a* and *b* are examples of *A* and *B*, then *a* must cause *b*. On deductive assessment of theory see, e.g., Hempel's discussion of "theoretical support" for theories in his *Philosophy of Natural Science*, pp. 38–40, and his related discussion of "deductive-nomological" explanations and "covering laws" on page 51 of the same work. The former are explana-

1. *Experimentation.* An investigator infers predictions from a theory. Then the investigator exposes only one of two equivalent groups to a stimulus. Are results congruent or incongruent with the predictions? Congruence of prediction and result corroborates the theory, incongruence infirms it.

2. *Observation.* An investigator infers predictions from a theory. Then the investigator passively observes the data without imposing an external stimulus on the situation and asks if observations are congruent with predictions.[30]

Predictions frame observations we expect to make if our theory is valid. They define expectations about the incidence, sequence, location, and structure of phenomena.[31] For instance, we can always predict that values on the independent and dependent variables of valid theories should covary across time and space, other things being equal. Values on intervening variables that form the theory's explanation should also covary with the independent variable across time and space. Variance on the independent vari-

---

tions that operate by deduction from general laws, the latter are general laws from which specific explanations are deduced.

Most "commonsense" explanations are theories we accept because they are supported by deductions of this sort; however, a deductive evaluation is not a test of a theory. Rather, it applies a previously tested law to a new situation.

30. Observation research designs are also called "quasi-experimental." See Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Boston: Houghton Mifflin, 1963), p. 34.

31. I use "prediction" to define expectations about the occurrence of phenomena in both the past and the future if a theory is valid. Others call these expectations the "observable implications" or the "test implications" of theory. King, Keohane, and Verba, *Designing Social Inquiry*, pp. 28–29 and passim; Hempel, *Philosophy of Natural Science*, pp. 7, 30. Still others use "postdiction" to refer to expectations about what the historical record will reveal, reserving "prediction" for expectations about the future.

We use predictions to design tests for hypotheses, but predictions are also hypotheses themselves. They frame phenomena that the independent variable should cause if the hypothesis operates. These phenomena include observable aspects of the dependent variable or intervening variables and effects that these variables produce. Thus the distinction between a prediction and a hypothesis lies not in their nature but the use to which they are put.

able should precede in time related variance on the dependent variable. If a social theory is being tested, actors should speak and act in a manner fitting the theory's logic (for example, if "commercial competition causes war," elites deciding for war should voice commercial concerns as reasons for war).

Some hard sciences (chemistry, biology, physics) rely largely on experiments. Others (astronomy, geology, paleontology) rely largely on observation. In political science experiments are seldom feasible, with rare exceptions such as conflict simulations or psychology experiments. This leaves observation as our prime method of testing.

Two types of observational analysis are possible.:

1. *Large-n,* or "statistical," analysis.[32] A large number of cases—usually several dozen or more—is assembled and explored to see if variables covary as the theory predicts.

2. *Case-study* analysis. The analyst explores a small number of cases (as few as one) in detail, to see whether events unfold in the manner predicted and (if the subject involves human behavior) whether actors speak and act as the theory predicts.[33]

Which method—experiment, large-*n*, or case study—is best? We should favor the method that allows the most strong tests. (I discuss strong tests later in this chapter.) More tests are better than fewer; strong tests are better than weak; many strong tests are best, as are methods that allow them. The structure of available data decides which method is strongest for testing a given theory.

32. Primers on large-*n* analysis include Babbie, *Practice of Social Research;* Shively, *Craft of Political Research;* William G. Cochran, *Planning and Analysis of Observational Studies* (New York: Wiley, 1983); Edward S. Balian, *How to Design, Analyze, and Write Doctoral or Masters Research,* 2d ed. (Lanham, Md.: University Press of America, 1988); Edward R. Tufte, *Data Analysis for Politics and Policy* (Englewood Cliffs, N.J.: Prentice-Hall, 1974); D. G. Rees, *Essential Statistics;* George W. Snedecor and William G. Cochran, *Statistical Methods* (Ames: Iowa State University Press, 1989); and David Freedman et al., *Statistics,* 2d ed. (New York: Norton, 1991).

33. Landmark writings on the case-study method are listed in note 1 to Chapter 2.

Most theories of war are best tested by case-study methods because the international historical record of prewar politics and diplomacy, which serves as our data, usually lends itself better to deep study of a few cases than to exploration of many cases. A few cases are recorded in great depth (the two World Wars) but the historical record deteriorates sharply as we move beyond the fifteenth or twentieth case. As a result case studies often allow more and stronger tests than large-*n* methods. Conversely, large-*n* methods are relatively more effective for testing theories of American electoral politics because very large numbers of cases (of elections, or of interviewed voters) are well recorded. Case studies can be strong tools for exploring American politics, however, especially if in-depth case studies yield important data that is otherwise inaccessible,[34] and large-*n* analysis can be a strong method for exploring international politics if relevant test data is recorded for many cases (see, for example, the many good large-*n* tests of democratic peace theory.)[35] Experimentation is the least valuable approach because experiments are seldom feasible in political science.

Strong vs. Weak Tests; Predictions and Tests

Strong tests are preferred because they convey more information and carry more weight than weak tests.[36]

34. Examples include Richard E. Fenno, *Home Style: House Members in Their Districts* (New York: HarperCollins, 1978), and Fiorina, *Congress: Keystone of the Washington Establishment.*

35. For example, Steve Chan, "Mirror, Mirror on the Wall . . . Are the Freer Countries More Pacific?" *Journal of Conflict Resolution* 28 (December 1984): 617–48; Erich Weede, "Democracy and War Involvement," ibid., pp. 649–64; and Zeev Maoz and Bruce Russett, "Normative and Structural Causes of Democratic Peace, 1946–1986," *American Political Science Review* 87 (September 1993): 624–38.

36. Discussions of strong tests include Eckstein, "Case Study and Theory," pp. 113–31, discussing what he terms "crucial-case studies" (his term for cases supplying strong tests), and Arthur L. Stinchcombe, *Constructing Social Theories* (New York: Harcourt, Brace & World, 1968), pp. 20–22.

A strong test is one whose outcome is unlikely to result from any factor except the operation or failure of the theory. Strong tests evaluate predictions that are *certain* and *unique.* A *certain* prediction is an unequivocal forecast. The more certain the prediction, the stronger the test. The most certain predictions are deterministic forecasts of outcomes that must inexorably occur if the theory is valid. If the prediction fails, the theory fails, since failure can arise only from the theory's nonoperation. A *unique* prediction is a forecast not made by other known theories. The more unique the prediction, the stronger the test. The most unique predictions forecast outcomes that could have no plausible cause except the theory's action. If the prediction succeeds, the theory is strongly corroborated because other explanations for the test outcome are few and implausible.

Certainty and uniqueness are both matters of degree. Predictions fall anywhere on a scale from zero to perfect on both dimensions. Tests of predictions that are highly certain and highly unique are strongest, since they provide decisive positive and negative evidence. As the degree of certitude or uniqueness falls, the strength of the test also falls. Tests of predictions that have little certitude or uniqueness are weakest, and are worthless if the tested prediction has no certitude or uniqueness.

We can distinguish four types of tests, differing by their combinations of strength and weakness:

1. *Hoop tests.* Predictions of high certitude and no uniqueness provide decisive negative tests: a flunked test kills a theory or explanation, but a passed test gives it little support. For example: "Was the accused in the state on the day of the murder?" If not, he is innocent, but showing that he was in town does not prove him guilty. To remain viable the theory must jump through the hoop this test presents, but passage of the test still leaves the theory in limbo.

2. *Smoking-gun tests.* Predictions of high uniqueness and no certitude provide decisive positive tests: passage strongly corrobo-

rates the explanation, but a flunk infirms it very little. For example, a smoking gun seen in a suspect's hand moments after a shooting is quite conclusive proof of guilt, but a suspect not seen with a smoking gun is not proven innocent. An explanation passing a "smoking-gun" test of this sort is strongly corroborated, but little doubt is cast on an explanation that fails it.

3. *Doubly-decisive tests.* Predictions of high uniqueness and high certitude provide tests that are decisive both ways: passage strongly corroborates an explanation, a flunk kills it. If a bank security camera records the faces of bank robbers, its film is decisive both ways—it proves suspects guilty or innocent. Such a test combines a "hoop test" and "smoking-gun" test in a single study. Such tests convey the most information (one test settles the matter) but are rare.

4. *Straw-in-the-wind tests.* Most predictions have low uniqueness and low certitude, and hence provide tests that are indecisive both ways: passed and flunked tests are both "straws in the wind." Such test results can weigh in the total balance of evidence but are themselves indecisive. Thus many explanations for historical events make probabilistic predictions ("If Hitler ordered the Holocaust, we should probably find some written record of his orders")[37], whose failure may simply reflect the downside probabilities. We learn something by testing such straw-in-the-wind predictions, but such tests are never decisive by themselves.[38] Unfortunately, this describes the predictions we usually work with.

Interpretive disputes often arise from disputes over what outcomes theories predict. Does Realism make predictions that were

37. In fact there is no written record of an order from Hitler mandating the Holocaust, yet historians agree that Hitler did order it. A discussion is Sebastian Haffner, *The Meaning of Hitler,* trans. Ewald Osers (Cambridge: Harvard University Press, 1979), pp. 133, 138–43.

38. These last four terms—"hoop test," "smoking-gun test," "doubly-decisive test," and "straw-in-the-wind test"—are my effort to fill gaps in the lexicon.

contradicted by the end of the cold war? Some scholars say yes, others say no. Such disagreements can be narrowed if theories are clearly framed to begin with (since vague theoretical statements leave more room for divergent predictions) and if tested predictions are explained and justified.

Interpretive disputes also arise from quarrels over the uniqueness and certitude of predictions. Is the prediction unique? That is, do other theories or explanations predict the same result? If so, a passed test is less impressive. The Fischer school of historians argues that the December 8, 1912, German "war council," a sinister meeting between Kaiser Wilhelm II and his military leaders (uncovered only in the 1960s), signaled a plot among the German elite to instigate a major war.[39] Some critics answer that the Kaiser's mercurial personality explains his bellicose talk at that meeting—he often blew off steam by saying things he did not mean. In short, they point to a competing explanation for events that some Fischerites claimed was a "smoking gun" for their elite-plot theory of the war. The question then rides on the plausibility of this competing explanation.

Is the prediction certain, in other words, is it unequivocal? If not, flunked tests are less damaging. Some historians argue that the Spanish-American war of 1898 arose from a conspiracy of empire-seeking U.S. leaders who hoped to seize the Philippines from Spain. The absence of any mention of such a conspiracy in these leaders' diaries and private letters or in official archives convinces others that there was none. In this view the conspiracy theory predicts with high certainty that mention of a conspiracy should be found in these records. Conspiracy theorists answer

39. On the "war council" see Imanuel Geiss, *German Foreign Policy, 1871–1914* (Boston: Routledge & Kegan Paul, 1976), pp. 142–45, 206–7. Good friendly surveys of the Fischer school's arguments are ibid., and John A. Moses, *The Politics of Illusion: The Fischer Controversy in German Historiography* (London: George Prior, 1975). More critical is John W. Langdon, *July 1914: The Long Debate, 1918–1990* (New York: Berg, 1990), pp. 66–129.

that good conspirators hide their conspiracies, often leaving no records. The conspiracy theory is still alive, they argue, because the theory predicts only weakly that conspirators should record their conspiracy, hence the lack of such a record is a mere "straw in the wind" that infirms the theory only weakly. The question hinges not on the evidence but on divergent estimates of the certitude of the theory's prediction that a conspiracy would leave a visible record.

This discussion highlights the need to discuss the uniqueness and certitude of tested predictions when interpreting evidence. All evidence is not equal because the predictions they test are not equally unique or certain. Hence authors should comment on the uniqueness and certitude of their predictions.

Strong tests are preferred to weak tests, but tests can also be hyper-strong, or unfair to the theory. For example, one can perform tests under conditions where countervailing forces are present that counteract its predicted action. Passage of such tests is impressive because it shows the theory's cause has large importance, that is, high impact. But a valid theory may flunk such tests because a countervailing factor masks its action. Such a test misleads by recording a false negative—unless the investigator, mindful of the test's bias, gives the theory bonus points for the extra hardship it faces.

Another form of hyper-strong test evaluates theories under circumstances that lack the antecedent conditions they require to operate. Again the theory is unlikely to pass, and we are impressed if it does. Passage suggests that the theory has wider explanatory range than previously believed. Such tests are not fair measures of a theory's basic validity, however, since they assess it against claims that it does not make.[40]

40. Advocates of testing theories against "least-likely" cases—cases that ought to invalidate theories if any cases can be expected to do so—recommend a hyper-strong test of this sort if the case they recommend is least-likely because it lacks conditions needed for the theory to operate. A flunked test then tells us that the

Helpful Hints for Testing Theories

Theory-testers should follow these injunctions:

1. Test as many of a theory's hypotheses as possible. Testing only a subset of a theory's hypotheses is bad practice because it leaves the theory partly tested. A theory is fully tested by testing all its parts.

The number of testable hypotheses exceeds the number of links in a theory. Consider the theory:

$$A \rightarrow q \rightarrow r \rightarrow B$$

A complete test would evaluate the theory's prime hypothesis ($A \rightarrow B$), the theory's explanatory hypotheses ($A \rightarrow q$, $q \rightarrow r$, and $r \rightarrow B$), and their hybrid combinations ($A \rightarrow r$ and $q \rightarrow B$). Thus a three-link theory comprises a total of six testable hypotheses. An analyst should explore them all, if time and energy permit.

2. Infer and test as many predictions of each hypothesis as possible. Most hypotheses make several testable predictions, so don't be quickly content to rest with one. To find more, consider what variance the hypothesis predicts across both time and space (that is, across regions, groups, institutions, or individuals). Consider also what decision process (if any) it predicts, and what specific individual speech and action it predicts.

Predictions frame observations you expect to make if the theory is valid. They define expectations about the incidence, sequence, location, and structure of phenomena. Avoid framing tautological predictions that forecast simply that we expect to observe the theory in operation ("If the theory is valid, I predict we will observe its cause causing its effect"). Thus the hypothesis that

---

theory will not operate if its antecedent conditions are absent, but it tells us nothing about the theory's validity when these conditions are met. Such tests are useful and appropriate if the scope of a theory's application is the main question, but are inappropriate if the validity of the theory is the question at issue. Discussing least-likely cases is Eckstein, "Case Study and Theory," p. 118.

"democracy causes peace" yields the following tautological prediction: "We should observe democracy causing peace." A nontautological prediction would be: "We should observe that democratic states are involved in fewer wars than authoritarian states."

3. Explain and defend the predictions you infer from your theory. As I noted earlier, scientific controversies often stem from disputes over which predictions can be fairly inferred from a theory and which cannot be. We then see scientists agree on the data but differ over their interpretation because they disagree on what the tested theories predict. Theorists can minimize such disputes by fully explaining and defending their predictions.

Predictions can be either general (the theorist predicts a broad pattern) or specific (the theorist predicts discrete facts or other single observations). General predictions are inferred from, and are used to test, general hypotheses ("If windows of opportunity and vulnerability drive states to war, states in relative decline should launch more than their share of wars"). Specific predictions are inferred from, and are used to test, both general hypotheses ("If windows of opportunity and vulnerability drive states to war, we should see Japan behave more aggressively as a window of opportunity opened in its favor in 1941") and specific explanations ("If a window of opportunity drove Japan to war in 1941 we should find records of Japanese decision makers citing a closing window as reason for war").

4. Select data that represent, as accurately as possible, the domain of the test. When using large-*n* test methods, select data that represent the universe defined by tested hypotheses. When using case-study methods, select data that represent conditions in the cases studied. Even data that represent the domain of the test only crudely can be useful.[41] Still, the more accurate the representation,

41. John J. Mearsheimer, "Assessing the Conventional Balance: The 3:1 Rule and Its Critics," *International Security* 13 (Spring 1989): 56–62, argues for and illus-

the better. Choosing evidence selectively—that is, favoring evidence that supports your hypothesis over disconfirming counterevidence—is disallowed, since such a practice violates the principle of accurate representation.

This rule is almost a platitude, but older political science literature (I am thinking of works in international relations) often broke it by "arguing by example." Examples are useful to illustrate deductive theories but only become evidence if they represent (even crudely) the complete relevant data base, and/or they are presented in enough detail to comprise a single case study.

5. Consider and evaluate the possibility that an observed relationship between two variables is not causal but rather results from the effect of a third variable.[42] Two variables may covary because one causes the other, or because a third variable causes both. For example, monthly sales of mittens and snow blowers correlate closely in the northern United States, but neither causes the other. Instead, winter weather causes both. We should consider or introduce controls on the effects of such third variables before concluding that correlation between variables indicates causation between them.

6. When interpreting results, judge each theory on its own merits.

If you flunk (or pass) a theory, do not assume a priori that the same verdict applies to similar theories. Each theory in a theory family (such as the neoclassical family of economic theories, the Marxist family of theories of imperialism, the Realist family of theories of international relations, and so on) should be judged on its own. The strengths and weaknesses of other theories in the family should not be ascribed to it unless both theories are vari-

---

trates the utility of "rule of thumb" tests using data that not selected for its representativeness.

42. A discussion is Babbie, *Practice of Social Research*, pp. 396–409.

ants of the same more general theory and your test has refuted or corroborated that general theory.

If you flunk (or pass) one hypothesis in a multihypothesis theory, this says nothing about the validity of other hypotheses in the theory. Some may be false and others true. You should test each separately.

Consider whether you can repair flunked theories before discarding them. Flunked theories often contain valid hypotheses. Perhaps they can be salvaged and incorporated into a new theory.

7. We can repair theories by replacing disconfirmed hypotheses with new explanatory hypotheses proposing a different intervening causal process or by narrowing the scope of the theory's claims. We narrow a theory's claims by adding new antecedent conditions (condition variables, or CVs), so the theory no longer claims to govern the cases comprised in the flunked test. This allows us to set aside the flunked test. The theory is now more modest but passes its tests.

8. We can test theories against the null hypothesis (the test asks, "Does this theory have *any* explanatory power?") or against each other (the tests asks, "Does this theory have *more* or *less* explanatory power than competing theories?").[43] Both test formats are useful but should not be confused. Theories that pass all their tests against the null should not be named the leading theory without

43. Imre Lakatos likewise distinguishes "a two-cornered fight between theory and experiment" and "three-cornered fights between rival theories and experiment." His "two-cornered fights" are tests against the null hypothesis (the hypothesis of no causal relationship); his "three-cornered fights" include a test against the null and a theory-against-theory test. Imre Lakatos, "Falsification and the Methodology of Scientific Research Programmes," in Imre Lakatos and Alan Musgrave, eds., *Criticism and the Growth of Knowledge* (Cambridge: Cambridge University Press, 1970), p. 115. Works formatted as two-cornered fights include many studies on democratic peace theory, for instance, Chan, "Mirror, Mirror on the Wall," and Weede, "Democracy and War Involvement." A study formatted as a three-cornered fight is Barry R. Posen, *The Sources of Military Doctrine: Britain, France, and Germany Between the World Wars* (Ithaca, N.Y.: Cornell University Press, 1984). For more on the topic see Hempel's discussion of "crucial tests" in his *Philosophy of Natural Science*, pp. 25–28.

further investigation; they can still lose contests against competing theories. Conversely, theories that lose contests against competitors should not be dismissed altogether. They may still have some explanatory power, and theories with explanatory power are valuable even if other theories have more.

9. One tests a theory by asking if the empirical evidence confirms the theory's predictions, not by asking how many cases the theory can explain. A theory may explain few cases because its causal phenomenon is rare or because it requires special hothouse conditions to operate, but can still operate strongly when these conditions are present. Such a theory explains few cases but is nevertheless valid.

The number of cases a theory explains does shed light on its utility: the more cases the theory explains, the more useful the theory, other things being equal. Still, even theories that explain very few cases are valuable if these cases are important and the theory explains them well.

10. One does not test a theory by assessing the validity of its assumptions (the assumed values on its CVs). A test asks: "Does the theory operate if the conditions that it claims to require for its operation are present?" Framed this way, a test axiomatically assumes assumptions are true. Tests under conditions that violate the theory's assumptions are unfair, and theories should not be rejected because they flunk such tests.

The validity of a theory's assumptions does affect its utility, however. Assumptions that never hold give rise to theories that operate only in an imaginary world and thus cannot explain reality or generate policy prescriptions.[44] The most useful theories are

44. For a different view see Friedman, *Essays in Positive Economics*, pp. 14–23: "In general, the more significant the theory, the more unrealistic the assumptions" (p. 14). Friedman's claim stems from his exclusive focus on the ability of theories to accurately predict outcomes (the values of dependent variables). He is uninterested in the validity of the inner workings of theories, including their explanations as well as their assumptions. This unconcern is appropriate if knowledge

those whose assumptions match reality in at least some important cases.

How Can Specific Events Be Explained?

Ideas framing cause and effect come in two broad types: theories and specific explanations. Theories are cast in general terms and could apply to more than one case ("Expansionism causes war," or "Impacts by extraterrestrial objects cause mass extinctions"). Specific explanations explain discrete events—particular wars, interventions, empires, revolutions, or other single occurrences ("German expansionism caused World War II," or "An asteroid impact caused the extinction of the dinosaurs"). I have covered the framing and testing of theories above, but how should we evaluate specific explanations?[45] We should ask four questions:

1. Does the explanation exemplify a valid general theory (that is to say, a covering law)?[46] To assess the hypothesis that *A* caused *b* in a specific instance, we first assess the general form of the hypothesis ("*A* causes *B*"). If *A* does not cause *B*, we can rule out all explanations of specific instances of *B* that assert that examples of *A* were the cause, including the hypothesis that *A* caused *b* in this case.

---

about the nature of the theory's inner workings is not useful, but this is seldom the case in the study of politics.

45. The role of theories in historical explanation has long been debated by historians and philosophers of social science. My remarks here follow Hempel, "Function of General Laws in History," the landmark work in the debate. For criticisms and other reactions see Martin and McIntyre, *Readings in the Philosophy of Social Science*, pp. 55–156. A recent discussion is Clayton Roberts, *The Logic of Historical Explanation* (University Park: Pennsylvania State University Press, 1996). See also Eckstein, "Case Study and Theory," pp. 99–104, who discusses "disciplined-configurative" case studies, that is, case studies that aim to explain the case by use of general theories.

46. A general theory from which a specific explanation is deduced is the "covering law" for the explanation. See Hempel, *Philosophy of Natural Science*, p. 51.

We assess the argument that "the rooster's crows caused today's sunrise" by asking whether, in general, roosters cause sunrises by their crowing. If the hypothesis that "rooster crows cause sunrises" has been tested and flunked, we can infer that the rooster's crow cannot explain today's sunrise. The explanation fails because the covering law is false.

Generalized specific explanations are preferred to nongeneralized specific explanations because we can measure the conformity of the former but not the latter to their covering laws. (The latter leave us with no identified covering laws to evaluate.) Nongeneralized specific explanations must be recast as generalized specific explanations before we can measure this conformity.

2. Is the covering law's causal phenomenon present in the case we seek to explain? A specific explanation is plausible only if the value on the independent variable of the general theory on which the explanation rests is greater than zero. Even if *A* is a confirmed cause of *B*, it cannot explain instances of *B* that occur when *A* is absent.

Even if economic depressions cause war, they cannot explain wars that occur in periods of prosperity. Even if capitalism causes imperialism it cannot explain communist or precapitalist empires. Asteroid impacts may cause extinctions, but cannot explain extinctions that occurred in the absence of an impact.

3. Are the covering law's antecedent conditions met in the case? Theories cannot explain outcomes in cases that omit their necessary antecedent conditions. Dog bites spread rabies if the dog is rabid; bites by a nonrabid dog cannot explain a rabies case.

4. Are the covering law's intervening phenomena observed in the case? Phenomena that link the covering law's posited cause and effect should be evident and appear in appropriate times and places. Thus if an asteroid impact killed the dinosaurs 65 million years ago, we should find evidence of the catastrophic killing process that an impact would unleash. For example, some scien-

tists theorize that an impact would kill by spraying the globe with molten rock, triggering forest fires that would darken the skies with smoke, shut out sunlight, and freeze the earth. If so, we should find the soot from these fires in 65-million-year-old sediment worldwide. We should also find evidence of a very large (continent-sized or even global) molten rock shower and a very abrupt dying of species.[47]

This fourth step is necessary because the first three steps are not definitive. If we omit step 4, it remains possible that the covering law that supports our explanation is probabilistic and the case at hand is among those where it did not operate.[48] We also should test the explanation's within-case predictions as a hedge against the possibility that our faith in the covering law is misplaced, and that the "law" is in fact false. For these two reasons, the better the details of the case conform to the detailed within-case predictions of the explanation, the stronger the inference that the explanation explains the case.[49]

47. In fact the sedimentary record laid down at the time of the dinosaurs' demise confirms these predictions. Walter Alvarez and Frank Asaro, "An Extraterrestrial Impact," *Scientific American*, October 1990, pp. 79–82.

The debate over the dinosaur extinction nicely illustrates the inference and framing of clear predictions from specific explanations. On the impact theory see Alvarez and Asaro, "Extraterrestrial Impact"; Vincent Courtillot, "A Volcanic Eruption," *Scientific American*, October 1990, pp. 85–92; and William J. Broad, "New Theory Would Reconcile Views on Dinosaurs' Demise," *New York Times*, December 27, 1994, p. C1.

48. The cause of probabilism in probabilistic causal laws usually lies in variance in the values of antecedent conditions that we have not yet identified. By identifying these conditions and including them in our theory we make its law less probabilistic and more deterministic.

49. Less convinced of the need for this last step is Hempel, "Function of General Laws in History," who rests with the first three steps and omits the fourth. Hempel assumes that his covering laws are deterministic (not probabilistic) and are well proven. Most social science laws are probabilistic, however, and most are poorly established. Hence deducing the validity of a specific explanation from the first three steps alone is unreliable, and we should also seek empirical verification that the explanation's causal process in fact occurred before reaching final conclusions.

Analysts are allowed to infer the covering law that underlies the specific explanation of a given event from the event itself. The details of the event suggest a specific explanation; the analyst then frames that explanation in general terms that allow tests against a broader database; the explanation passes these tests; and the analyst then reapplies the theory to the specific case. Thus the testing of the general theory and the explaining of a specific case can be done together and can support each other.

Methodology Myths

Philosophers of social science offer many specious injunctions that can best be ignored. The following are among them:

1. "Evidence infirming theories transcends in importance evidence confirming theories." Karl Popper and other falsificationists argue that "theories are not verifiable," only falsifiable,[50] and that tests infirming a theory are far more significant than tests confirming it.[51] Their first claim is narrowly correct, their second is not. Theories cannot be proved absolutely because we cannot imagine and test every prediction they make, and the possibility always remains that an unimagined prediction will fail. By contrast, infirming tests can more decisively refute a theory. It does not follow that infirming tests transcend confirming tests, however. If a theory passes many strong tests but then flunks a test of a previously untested prediction, this usually means that the theory requires previously unidentified antecedent conditions to operate.

50. Karl R. Popper, *The Logic of Scientific Discovery* (London: Routledge, 1995), p. 252. A criticism of Popper and falsificationism is King, Keohane, and Verba, *Constructing Social Inquiry*, pp. 100–103.

51. In a friendly summary of falsificationism David Miller writes that to falsificationists "the passing of tests . . . makes not a jot of difference to the status of any hypothesis, though the failing of just one test may make a great deal of difference." David Miller, "Conjectural Knowledge: Popper's Solution of the Problem of Induction," in Paul Levinson, ed., *In Pursuit of Truth* (Atlantic Highlands, N.J.: Humanities Press, 1988), p. 22.

We react by reframing the theory to include the antecedent condition, thus narrowing the scope of the theory's claims to exclude the flunked test. In Popper's terms we now have a new theory; however, all the tests passed by the old theory also corroborate the new, leaving it in very strong shape at birth. Thus confirming tests tell us a great deal—about the old theory, about its repaired replacement, and about any later versions. Popper's contrary argument stems partly from his strange assumption that once theories are stated they are promptly accepted,[52] hence evidence in their favor is unimportant because it merely reinforces a preexisting belief in the theory. The opposite is more often true: most new ideas face hostile prejudice even after confirming evidence accumulates.[53]

2. "Theories cannot be falsified before their replacement emerges." Imre Lakatos claims that "there is no falsification [of theory] before the emergence of a better theory," and "falsification cannot precede the better theory."[54] This claim is too sweeping. It applies only to theories that fail some tests but retain some explanatory power. We should retain these theories until a stronger replacement arrives. But if testing shows that a theory has no explanatory power, we should reject it whether or not a replacement theory is at hand.[55] Many science programs—for example, medical research—advance by routinely testing theories against null hypotheses and rejecting those that fail, whether or not replacements are ready.

52. See King, Keohane, and Verba, *Designing Social Inquiry*, p. 100.

53. A famous development of this argument is Thomas S. Kuhn, *The Structure of Scientific Revolutions*, 2d enlarged edition (Chicago: University of Chicago Press, 1970).

54. Lakatos, "Falsification and the Methodology of Scientific Research Programmes," pp. 119, 122.

55. An early reader of this chapter suggested that Lakatos meant only that falsification of theories that retain some explanatory power cannot precede the better theory, following the argument I suggest here. That may be the case. Lakatos's arguments are well hidden in tortured prose that gives new meaning to the phrase "badly written," and no reading of such dreadful writing is ever certain or final.

Asking those who claim to refute theories or explanations to propose plausible replacements can serve as a check on premature claims of refutation. This can expose instances where the refuting investigator held the theory to a standard that their own explanation could not meet. This suggests in turn that the standard was too high, in other words, that the refuter misconstrued noise in the data as decisive falsifying evidence against the theory. However, finding merit in this exercise is a far cry from agreeing that theories cannot be falsified except by the greater success of competing theories. Surely we can know what is wrong before knowing what is right.

3. "The evidence that inspired a theory should not be reused to test it." This argument[56] is often attached to warnings not to test theories with the same cases from which they were inferred. It rests on a preference for blind testing.[57] The assumption is that data not used to infer a theory is less well known to an investigator than used data, hence the investigator using unused data is less tempted to sample the data selectively.

Blind testing is a useful check on dishonesty, but is not viable as a fixed rule. Its purpose is to prevent scholars from choosing corroborating tests while omitting infirming ones. But imposing blind-test rules on social science is in fact impossible because investigators nearly always know something about their data before they test their theories and thus often have a good idea what tests will show even if they exclude the data that inspired their ideas.

56. Raising this issue are Alexander L. George and Timothy J. McKeown, "Case Studies and Theories of Organizational Decision Making," in *Advances in Information Processing in Organizations* (Greenwich, Conn.: JAI Press, 1985), 2:38; David Collier, "The Comparative Method," in Ada W. Finifter, ed., *Political Science: The State of the Discipline*, 2d ed. (Washington, D.C.: American Political Science Association, 1993), p. 115; and King, Keohane, and Verba, *Designing Social Inquiry*, pp. 21–23, 46, 141, who note "the problem of using the same data to generate and test a theory . . ." (p. 23) and argue that "we should always try to . . . avoid using the same data to evaluate the theory that we used to develop it" (p. 46).

57. A discussion is Hempel, *Philosophy of Natural Science*, pp. 37–38.

Hence we need other barriers against test fudging.[58] Infusing social science professions with high standards of honesty is the best solution.

4. "Do not select cases on the dependent variable"—that is, do not select cases of what you seek to explain (for example, wars) without also choosing cases of the contrary (peace). Students of the case method often repeat this warning.[59] It is not valid. Selection on the dependent variable is appropriate under any of three common conditions:

a. If we can compare conditions in selected cases to a known average situation.[60] The average situation is often sufficiently well known not to require further descriptive study. If so, we can com-

58. Moreover, a blind-test requirement would generate a preposterous double standard in the right to use evidence: the same data would be forbidden as test material to some scholars (because they inferred the theory from it) while being allowed to others. How would this rule be administered? Who would record which scholars had used which data for theory-making, and hence were barred from reusing it for testing? Would we establish a central registry of hypotheses where theorists would record the origins of their ideas? How would we verify and penalize failure to accurately record hypotheses with this registry? How would we deal with the many scholars who are not really sure where their hypotheses come from?

59. See Barbara Geddes, "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Cases," *Political Analysis* 2 (1990): 131–50; also King, Keohane, and Verba, *Designing Social Inquiry*, pp. 108–9, 129–32, 137–38, 140–49. King et al. warn that "we will not learn anything about causal effects" from studies of cases selected without variation on the dependent variable; they declare that the need for such variation "seems so obvious that we would think it hardly needs to be mentioned"; and they conclude that research designs that lack such variation "are easy to deal with: avoid them!" (pp. 129–30). A criticism is Ronald Rogowski, "The Role of Scientific Theory and Anomaly in Social-Scientific Inference," *American Political Science Review* 89 (June 1995): 467–70. Rogowski notes that King, Keohane, and Verba's strictures point to a "needlessly inefficient path of social-scientific inquiry," and obedience to these strictures "may paralyze, rather than stimulate, scientific inquiry" (p. 470). On Geddes and King, Keohane, and Verba see also David Collier and James Mahoney, "Insights and Pitfalls: Selection Bias in Qualitative Research," *World Politics* 49 (October 1996): 56–91.

60. Thus Lijphart notes the "implicitly comparative" nature of some single-case studies. "Comparative Politics and the Comparative Method," pp. 692–93.

pare cases selected on the dependent variable to these known normal conditions. There is no need for full-dress case studies to provide sharper points of comparison.[61]

b. If the cases have large within-case variance on the study variable, permitting multiple within-case congruence procedures.

c. If cases are sufficiently data-rich to permit process tracing.[62]

These conditions allow test methods—comparison to average conditions, multiple within-case congruence procedures, and process tracing—that do not require comparison to specific external cases. When they are used, failure to select cases for explicit comparison raises no problems.

5. "Select for analysis theories that have concepts that are easy to measure." Some scholars recommend we focus on questions that are easy to answer.[63] This criterion is not without logic: study of the fundamentally unknowable is futile and should be avoided. However, the larger danger lies in pointlessly "looking under the light" when the object sought lies in darkness but could with effort be found. Large parts of social science have already diverted their focus from the important to the easily observed, thereby drifting into trivia.[64] Einstein's general theory of relativity proved hard to test. So should he have restrained himself from devising it? The structure of a scientific program is distorted when researchers shy from the logical next question because its answer will be hard to

61. Thus the erring scholars that Geddes identifies erred because they misconstrued the normal worldwide background levels of the key independent variables, e.g. intensity of labor repression, that they studied.

62. On congruence procedure and process tracing see the section "Testing Theories with Case Studies," in Chapter 2.

63. King, Keohane, and Verba warn that "we should choose observable, rather than unobservable, concepts wherever possible. Abstract, unobserved concepts such as utility, culture, intentions, motivations, identification, intelligence, or the national interest are often used in social science theories," but "they can be a hindrance to empirical evaluation of theories . . . unless they can be defined in a way such that they, or at least their implications, can be observed and measured." King, Keohane, and Verba, *Constructing Social Theories*, p. 109.

64. See, for example, the last several decades of the *American Political Science Review*.

find.[65] A better solution is to give bonus credit to scholars who take on the harder task of studying the less observable.

6. "Counterfactual analysis can expand the number of observations available for theory-testing." James Fearon suggests this argument.[66] Counterfactual statements cannot provide a substitute for empirical observations, however. They can clarify an explanation: "I claim x caused y; to clarify my claim, let me explain my image of a world absent x." They can also help analysts surface hypotheses buried in their own minds (see the section "How Can Theories Be Made?" in this chapter). But counterfactual statements are not data and cannot replace empirical data in theory-testing.

65. Moreover, tests that are difficult for the time being may become feasible as new tests are devised or new data emerge. Thus theories of the Kremlin's conduct under Stalin were hard to test before the Soviet collapse but later became more testable. This is another reason to keep hard questions on the agenda.
66. James D. Fearon, "Counterfactuals and Hypothesis Testing in Political Science," *World Politics* 43 (January 1991): 171 and passim.