

TABLE OF CONTENTS

Introduction

Memo 1: Hypotheses, Laws and Theories: A User's Guide

What is a Theory?

What is a Specific Explanation?

What is a Good Theory?

How Can Theories Be Made?

How Can Theories Be Tested?

Helpful Hints for Testing Theories

How Can Specific Events Be Explained?

Methodology Myths

Memo 2: What Are Case Studies? How Should They Be Performed?

Case Studies in Perspective

Testing Theories with Case Studies

Creating Theories with Case Studies

Inferring Antecedent Conditions with Case Studies

Testing Antecedent Conditions with Case Studies

Explaining Cases

Strong/Weak Tests: Predictions and Tests

Interpreting Contradictory Results

Case Selection Criteria

Memo 3: What is a Political Science Ph.D. Dissertation?

Theory-Proposing Dissertations

Theory-Testing Dissertations

Literature-Assessing or Stock-Taking Dissertations

Policy-Evaluative or Policy-Prescriptive Dissertations

Historical Explanatory Dissertations

Historical Evaluative Dissertations

Predictive Dissertations

Memo 4: Helpful Hints on Writing a Political Science Ph.D. Dissertation

Topic Selection

Organization

Your Dissertation Prospectus

Your Introductory Chapter

Your Concluding Chapter

Study Design and Presentation

Writing

Style

Vetting

Your Abstract

Dealing with your Thesis Committee

INTRODUCTION

Dealing with your Family, your Friends, and your Head 62
 How to Learn More about How to Write a Thesis 63
 Memo 5: The Ph.D. Dissertation Proposal 64
 Memo 6: How to Write a Paper 65
 Memo 7: Professional Ethics 68

TABLES

Table 1: Nine Case Selection Criteria: When Is Each Appropriate? 49

The seven memos in this working paper distill advice on methodology that I given, received or overheard over the years. The memos focus on issues that see get the most air time in classroom and hallway discussions occurring in my ge neighborhood. They began as how-to-do-it class handouts that I drafted to sav from further repeating oft-repeated advice. Any advice that I could remember g more than once qualified for inclusion, however mundane it was ("begin parag with topic sentences!") If a point needs saying it needs saying. At the same t omitted standard advice that seldom needed giving.

Thus these memos are not definitive or exhaustive. All seven are basic pri designed as places to start. I make no effort to cover the methodological wate (Main omission: I have no memo on large-n study methods.) All express a pe viewpoint, and hardly claim to summarize or distill the views of the field. Early r drafts were designed less to offer conclusive answers than to spur students to de their own answers to the puzzles the memos address.

The views I express arise more from practicing methodology than studying it. reflect my experience as a student, colleague, teacher, and editor. I learned some i tant things from writings on philosophy of science and social science methods, have found the bulk of that writing abstruse, impractical, and unuseful. It was offe to invent my own answers than dig them out of the piles of written mud produc philosophers and methodologists, even when the answers existed somewhere i mud. (The reader may here correctly detect some exasperation with the methodolog erature. These memos should have been easier to write, and in fact should not needed writing: methodologists should have produced them long ago, leaving p like myself to get on with other things.)

These memos reflect my field of concentration (international relations/se affairs). They focus on concerns that are greater in the IR/security subfield (e.g case study method), and their footnotes and examples are IR-heavy. However, points discussed seem broadly relevant to the political science field, and readers in IR subfields may find them useful. My regrets in advance to these readers fr memos' parochial IR focus, which reflects their class-handout origins.

Several memos have rough spots.² This, too, reflects their origin: as unpubl memos I drafted them with little eye on esthetics. However, as the years passed n lished primers appeared to replace them, and others have found them useful. It : unconstructive to withhold papers that others can use because they have not yet polished to sublime perfection. Hence I am letting them go forth despite their blem

¹ Many fine primers on large-n methods exist already, so my absent memo should not be badly missed. I list pri large-n in footnote 31 of Memo 1, "Hypotheses, Laws and Theories: A User's Guide."

² E.g., readers will notice that sections II, III, IV, V, and IX in Memo 2 are somewhat repetitive, and sections III an Memo 1 overlap with sections VI and VII in Memo 2.

Memo 1, "Hypotheses, Laws and Theories: A User's Guide," began as a 3-page beginner's handout on scientific inference for an undergraduate class I taught at U.C. Davis many years ago. Finding no primer that explained how to frame, assess and apply theories I wrote my own guidelines. They grew and multiplied over the years, and reflect my own barefoot positivist, anti-obfuscationist viewpoint. I am unpersuaded by the view that the prime rules of scientific method should differ between hard science and social science. Science is science. I also believe the basic rules of science, often complexified, are actually few in number and can be plainly stated and briefly summarized.

The subjects of Memo 1--the basic tools and basic rules of scientific inference--are often skipped over in the methods texts. Much writing on social science methods assumes that readers already know what theories are, what good theories are, what elements theories contain, how theories should be expressed, what fundamental rules should be followed when testing or applying theories, etc. This memo stresses the elementary points others omit.

Memo 2, "What are Case Studies? How Should They Be Performed?", began as a graduate class handout drafted to escort an assignment to produce a short case study. Finding no short primer on how to do case studies, I wrote my own. I intended it as a starting point for students with no exposure to the case method.

Case study is the poor cousin among social science methods. The mainstream methodology literature pays vast attention to large-n methods while dismissing case methods with a wave. Many political science graduate programs teach large-n methods as the only technique: "methodology" classes cover large-n methods (or large-n and rational choice) as if these were all there is. Case study methods are seldom taught, and almost never in a class of their own. (Exceptions include classes on the case method taught by Stephen Walt and John Mearsheimer at the U of Chicago, by Scott Sagan at Stanford, by Peter Liberman at Tulane, and in the past by Ted Hopf at the University of Michigan.)

I regard large-n and case study methods as essentially co-equal. Each has strengths and weaknesses. Sometimes one is the stronger method, sometimes the other. Hence the imbalanced attention they receive should be righted.³ However, the fault for this imbalance lies partly with case study practitioners themselves. They have produced no how-to cookbooks that show novices the ropes. With no cookbook distilling the method, others are bound to neglect it. I wrote this memo as a first cut toward such a cookbook.

Memo 3, "What is a Political Science Ph.D. Dissertation?", reflects my view that we often define the boundaries of the field too narrowly. A wider range of thesis topics and formats should be considered fair game. Specifically, political science field culture is biased toward the creation and testing of theory over other work, including the application of theory and the stock-taking of literature. But making and testing theories are not the only games in town. Applying theories to evaluate past and present policies and

to solve historical puzzles is also worth doing. If everyone makes and tests it no one ever uses them, then what are they for? Stock-taking work is also invaluable, as our literature expands to the point where no one can take stock on their own.

Moreover, theory making and theory testing can be tall orders for scholarship beginning of their careers. Grand theorizing takes time to learn, and applying or taking stock can be a more feasible way to start. Theory-application and writing both require good facility with theories, and both allow wide latitude to do that facility with less risk of utter failure. Hence theses of this genre should be organized as respectable social science, and should be considered as alternative grand-theory options look daunting.

Memo 3 also reflects my view that political science should embrace the historical explanation among its missions. Historians should not be left alone with their theories. Many historians are leery of generalizations, hence of the use of general theories are essential to historical explanation. Many are averse to explicit evaluation instead preferring to "let the facts speak for themselves." Many are averse to evaluative history. Political science should step in to fill the explanatory and gaps that are left by these quirks in historian culture.

Memos 4 and 5, "Helpful Hints on Writing a Political Science Ph.D. Dissertation" and "The Ph.D. Dissertation Proposal," distill craft advice that I have given and colleagues over the years, and that others have given to me. They focus on presentation and on broader questions of academic strategy and tactics, writing the (usually more important) questions of research design. In part they are time editing *International Security*, and the many discussions I had about its presentation with IS readers and authors.

Memo 6, "How to Write a Paper," is a class-handout memo on writing my undergraduate students along with all paper assignments. What should a look like? These are my marching orders. Everyone has their own taste in and this memo distills mine.

I don't fancy myself a great writer or writing teacher. However, teaching graduates how to write is among the most important duties of the college teacher.⁴ To teach writing well a teacher needs a written advisory for students to write. Finding no short primer that did this well I wrote my own, although special qualifications on the matter. I include it here with the thought that others may use it (or adapt it) as their own class how-to writing handout.

Memo 7, "Professional Ethics," is somewhat off the narrow methodology of the others, but it does touch methodology in a large sense by asking: how work together as a community? It reflects my feeling that the social sciences formal discussion of professional ethics. Social science operates largely

³The general dismissal of the case method is especially harmful to the study of international relations, because the structure of IR data often better lends itself to the case method than the large-n method.

⁴And if so, it does seem an oddity that most graduate social science programs teach future college teachers writing and nothing about the teaching of writing. Do we need a change of graduate training norms in

accountability to others. All institutions and professions that face weak accountability need inner ethical rudders that define their obligations in order to say on course. Otherwise they risk straying into parasitic disutility. Social science is no exception.

For educating me on many matters discussed here I thank Robert Arseneau, with whom I discussed many of these issues while he taught PS3 at U.C. Berkeley. For comments on these memos and these issues I also thank Steve Ansolabehere, Bob Art, Tom Christensen, Charlie Glaser, Chaim Kaufmann, Peter Liberman, John Mearshimer, Scott Sagan, Jack Snyder, Marc Trachtenberg, Steve Walt, Sandy Weiner, and David Woodruff. I also thank *en masse* the many teachers, students and colleagues who have given me comments on my own work through the years. Much of the advice offered here is recycled advice that they once gave me.

MEMO 1:

HYPOTHESES, LAWS AND THEORIES: A USER'S GUIDE

1. WHAT IS A THEORY?

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

Theories are abstract general statements that describe and explain the causes or effects of classes of phenomena. They are composed of causal laws or hypotheses, explanant and antecedent conditions. Explanations are also composed of causal laws or hypotheses which are in turn composed of dependent and independent variables. Twelve definitions bear mention:

"Law":

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

Laws can also be causal ("A causes 'B'") or spurious ("A" and "B" are correlated but do not cause each other").² Our prime search is for causal laws. We explore the possibility that laws are spurious mainly to rule it out, so we can rule possibility that observed laws are causal.³

¹ Most posit that theories explain phenomena and leave it at that. The elements of an explanation are not detailed. Example, Brian Fay and I. Donald Moon, "What Would an Adequate Philosophy of Social Science Look Like?" in Martin and Lee C. McIntyre, eds., *Readings in the Philosophy of Social Science* (Cambridge: MIT Press, 1999), p. 35 at 26: a social theory is a "systematic, unified explanation of a diverse range of social phenomena." Like Babbie, *The Practice of Social Research*, 7th ed. (Belmont, Calif.: Wadsworth, 1995), p. 40: "A theory is a explanation for the observations that relate to a particular aspect of life." See also Kenneth Waltz, quoted in below. Each leaves the components of an explanation unspecified.

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

³ Generic laws (which might be causal or spurious) should be stated in associative language ("if 'A', then 'B'"; or "A", the greater "B"). Causal laws can also be framed with causal language ("A" causes "B")

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-determined causation ("A" causes "B" and "B" causes "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"). To complete it may also investigate whether reciprocal causation or self-determined 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 non-causal, i.e., positing a spurious relationship ("I surmise that 'A' and 'B' are caused by 'C'; hence 'A' and 'B' are correlated but do not cause each other"). However, the term "hypothesis" is widely used to refer only to causal hypotheses, and I use it with that narrower meaning here.

"Theory":

A causal law ("I have established that 'A' causes 'B'") or hypothesis ("I surmise that 'A' causes 'B'") and

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 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, and causes more growth 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.")

Like an explanation, an antecedent condition can be re-stated 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."

⁴This 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." Using "hypothesis" more broadly, to include conjectures about facts as well as relationships and hence including descriptive conjectures as hypotheses, is Carl G. Hempel, *Philosophy of Natural Science* (Englewood Cliffs, N.J.: Prentice-Hall, 1966), p. 19. I use "propositions" for what Hempel calls "hypotheses"; propositions can be hypotheses or descriptive conjectures. Also using "hypothesis" broadly, to include predictions inferred from hypotheses (below called the "predictions," "observable implications" or "test implications" of theory), is Babbie, *Practice of Social Research*, p. 49.

⁵The term is from Carl G. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press, 1965), pp. 246-247 and passim. The term "antecedent" merely means that the condition's presence precedes 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):

The phenomenon being caused.

"Independent Variable" (IV):

The phenomenon doing the causing. If "democracy causes peace" democracy is the independent, peace the dependent variable.

"Intervening Variable" (IntV):

The phenomena that form the theory's explanation. These are caused by the independent variable and cause the dependent variable.⁶

"Condition Variable" (CV):⁷

Antecedent conditions. Their values govern the size of the impact IVs or IntVs have on DVs and other IntVs.

"Study Variable" (SV):

A variable whose causes or effects we seek to discover with research. A project's study variable can be an IV, DV, IntV, or CV.

"Prime Hypothesis" (PH):

The over-arching hypothesis that frames the relationship between theory's independent and dependent variables.

"Explanatory Hypotheses" (EH):

The lesser-included hypotheses that comprise a theory's explanation.

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

⁶Whether a specific variable is dependent, independent, or intervening depends on its context and changes with a change in these statements:
"A' causes 'B'": A' is the independent variable.
"Q' causes 'A'": A' becomes the dependent variable.
"Q' causes 'A' causes 'B'": A' becomes an intervening variable.

⁷Condition variables are also known as "suppressor" variables, meaning that high values on these variables suppressular variance between independent and dependent variables. See Miller and Wilson, *Dictionary of Social Methods*, p. 110.
⁸These last four terms—"condition variable," "study variable," "prime hypothesis," and "explanatory hypotheses"—own nominations to fill word-gaps in the lexicon.

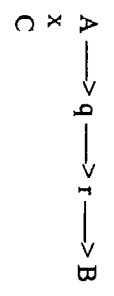
⁹For a different view see Kenneth N. Waltz, *Theory of International Politics* (Reading, Mass.: Addison-Wesley, 1979), p. 2, 5. To Waltz, theories are not "mere collections of laws" but are instead "statements that explain them" (p. 2, 5). Waltz's statements include "theoretical notions" which can take the form of concepts or assumptions. I prefer my 'a' statements because all explanations for social science laws that I find satisfying can be reduced to laws or hypotheses because all explanations also lack precision because it leaves the prime elements of an explanation unspecified 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 Snidal, "Rational Theory and Comparative Case Studies," *World Politics*, Vol. 41, No. 2 (January 1989), pp. 143-169 at 147: "A very general set of propositions from which others, including 'laws,' are derived." Their definition omits modal ideas that I call theories.
Nearer my usage is Carl Hempel: "Theories ... are bodies of systematically related hypotheses." Carl G. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press, 1965), pp. 246-247 and passim. The term "antecedent" merely means that the condition's presence precedes 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.

Theories can always be drawn with arrow diagrams, like this:



In this diagram 'A' is the theory's independent variable, 'B' is the dependent variable. 'q' and 'r' are intervening variables and comprise the theory's explanation. The proposal "A \longrightarrow B" is the theory's prime hypothesis, while the proposals that "A \longrightarrow q", "q \longrightarrow r", and "r \longrightarrow B" are its explanatory hypotheses.

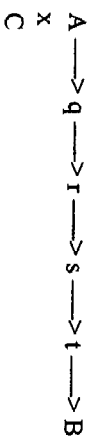
Condition variables can be added and defined with a "times" sign, "x".¹⁰ Here "C" is a condition variable: the impact of 'A' on 'q' is magnified by the presence of 'C' and reduced by 'C's absence.



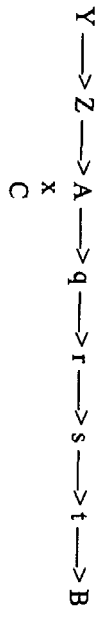
An example would be:



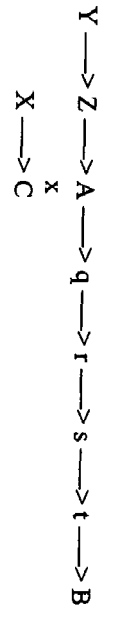
A theory's explanation can be displayed at any level of detail. Here the link between 'r' and 'B' has been elaborated to show explanatory variables 's' and 't'.



An explanation can be extended to define remote causes. Here remote causes of 'A' are detailed:

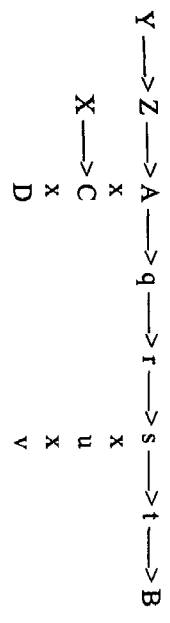


The causes of condition variables can be detailed, as here with the cause of 'C':

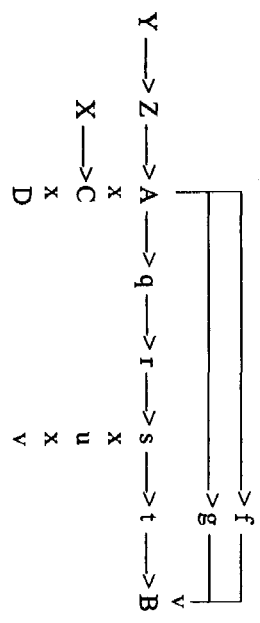


¹⁰The times sign is used loosely, to indicate that the CV magnifies the impact of the IV, but does not literally mean that the CV multiplies the impact of the IV (although it might.)

There is no limit to the number of antecedent conditions that might be framed. more conditions ('D', 'u', 'v') are specified.



More avenues of causation between causal and caused variables can be added. Here chains of causation between 'A' and 'B' (running through intervening variables 'f' and 'g' are added, to produce a three-chain theory:



A "theory" that cannot be arrow-diagrammed is not a theory and needs reform to become a theory. (Using this criteria much political science "theory" and "theoretical writing is not theory.)

II. WHAT IS A SPECIFIC EXPLANATION?

Explanations of specific events use theories and are framed like theories. A 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 mention:

"Specific explanation":

An explanation cast in specific terms that applies to a distinctive Like a theory it describes and explains cause and effect, but these es and effects are framed in singular terms. (Thus "expansionism es aggression, causing war" is a theory; "German expansionism causing German aggression, causing World War II" is a specific explanation The causes of the explained effects are no longer independent ables. Rather, they are concrete examples of high (or low) value independent variable. Specific explanations are also called "part explanations" (as opposed to a "general explanations.")

Specific explanations come in two types, with the second ("generalized specific explanation") the more useful type:

"Non-generalized 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 question "of what is Germany an example?")¹¹

"Generalized specific explanation": A specific explanation that identifies the theories that govern its operation ("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.")

The causal and caused phenomena that comprise specific explanations are not "variables" (since they are not framed as abstract phenomena whose values could vary). The following language, which avoids the term "variable," is more appropriate:

"Outcome Phenomenon" (OP): The phenomenon being caused.

"Causal Phenomenon" (CP): The phenomenon doing the causing.

"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.¹²

Specific explanations are arrow-diagrammed the same way theories are diagrammed:

A theory: Expansionism → Aggression → War

A generalized specific explanation: German expansionism → German aggression → World War II

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

¹²These last seven terms—"specific explanation," "non-generalized specific explanation," "generalized specific explanation," "outcome phenomenon," "causal phenomenon," "intervening phenomenon," and "antecedent phenomenon"—are my own suggested labels for these concepts. Others use "explanandum phenomenon" for the outcome phenomenon, and "explanans" for a generalized explanation and its components (the causal, intervening, and antecedent phenomena). E.g., Hempel, *Philosophy of Natural Science*, p. 50. (In Hempel's usage only generalized specific explanations comprise an explanans—non-generalized specific explanation do not.)

A non-generalized specific explanation: Germany → Outbreak of fighting on September 1, 1914 → World War II

III. WHAT IS A GOOD THEORY?

Eight prime attributes govern a theory's quality.

1. A good theory has *large explanatory power*. The theory's posited cause (independent variable) has a large effect on a wide range of phenomena under range of conditions. Three characteristics govern explanatory power:
 - a. **Size of impact:** does variance on the independent variable produce large variance on the dependent variable?¹³ If the IV has no impact the theoretical validity, hence no explanatory power. The greater the variance produced, the greater the theory's explanatory power.
 - b. **Explanatory range:** how many classes of phenomena are affected by, b explained by, the theory's independent variable? The wider the range of phenomena the greater the theory's explanatory power. Most social sciences have narrow range but a few gems explain many diverse domains.¹⁴
 - c. **Applicability:** how common is the theory's cause in the real world? How are antecedent conditions that activate its operation? The more prevalent and conditions of the theory, the greater its explanatory power.¹⁵ The present these causes and conditions in the past govern its power to explain historical current and future prevalence govern its power to explain present and future

¹³Impact size can be measured in "theoretical" or "dispersion" terms. A theoretical impact measurement as units of change in the dependent variable are caused by a unit change in the independent variable? (How many votes can a candidate gain by spending an additional campaign dollar on television ads?) A dispersion measure of the DV's total variance in a specific data set is caused by variance of this IV? (What percentage of the votes received by various congressional candidates is explained by variance in their television special advertising and Usine Regression (Beverly Hills: Sage, 1982), pp. 68-77. (Athen uses different language "importance" to refer to the impact of a variable.)

¹⁴Karl Deutsch used the terms "combinatorial richness" and "organizing power" for attributes similar explanatory range, with "combinatorial richness" expressing "the range of combinations or patterns that can be derived from a model, and "organizing power" defining the correspondence of the theory or model to phenomena in the world it was first used to explain. Karl Deutsch, *The Nerves of Government* (New York: Free Press, 1966), pp. 16-17. of social science theories with wide explanatory range include Mancur Olson's theory of public goods, Andrej's military-participation ratio (MPR) explanation for social stratification, and Stephen Walt's balance of power theory of alliances. See Mancur Olson, *The Logic of Collective Action* (Cambridge: Harvard University Press, 1965); Stephen M. Walt, *The Origins of Alliances* (Ithaca: Cornell University Press, 1987), pp. 17-33.

¹⁵Even causes that produce powerful effects can have little explanatory power if the causes themselves are rare, or if they require rare hot-house antecedent conditions to operate. Conversely, causes that produce great effects are often lethal, but they explain few deaths because they are scarce in the real world. The cause is hence it explains little. Sunburn is less lethal but explains more deaths (through skin cancer) because it is so common. Likewise, scuba diving is often lethal if hungry great white sharks are around, but scuba diving explains these deaths because divers avoid shark-infested waters. The cause is powerful under the right conditions (hungry sharks are rare), hence the cause explains few events. Sunburn explains more deaths because it is rare conditions to produce its harmful effects.

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

3. A good theory is "*satisfying*," i.e., it satisfies our curiosity. A theory is unsatisfying if we are left 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 his election loss: "I didn't get enough votes!" This is true but unsatisfying. We still want to know why he didn't get enough votes.

The farther 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. Otherwise we cannot tell what the theory predicts and so cannot perform tests.

A clearly-framed theory includes a full explanation 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."

5. A good theory includes a clear statement of the *antecedent conditions* that enable its operation and govern its impact. Otherwise 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. "Appeasing other states makes them more aggressive, causing war." This was true with Germany during 1938-1939, but sometimes the opposite is true—a firm stand makes others more aggressive, causing war. To avoid mistakes policymakers must know the antecedent conditions that decide if a firm stand makes others more or less aggressive. Hypotheses are more useful when their antecedent conditions are clearly framed.

6. A good theory is in principle *falsifiable*. Data that would falsify the theory can be defined (although it may not now be available.)¹⁵

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

Theories that make *omni-predictions* that are fulfilled by all observed events are non-falsifiable. Empirical tests cannot bolster or weaken such theories because all evidence is consistent with them. Religious theories of phenomena have this quality: they come as God's reward, disasters are God's punishment, cruelties are God's test of faith, and outcomes that elude these broad categories are God's mysteries. Many arguments share this *omni-predictional* trait.¹⁷

7. A good theory *explains important phenomena*: it answers questions that the wider world or it helps others answer such questions. Theories that answer questions are less useful even if they answer these questions well. (Much social theorizing has little real-world relevance and thus fails this test.)

8. 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 controllable by human action. Thus "capitalism causes inflation, 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 economic manipulation is more national military postures and doctrines. "Chauvinist history teaching in national school systems causes war" is even more useful, since the content of national education is more subject to control than national military policy.

A theory also gains prescriptive richness by identifying dangers that could, without being, be defeated or mitigated by timely countermeasures. Thus theories explaining causes of hurricanes provide no way to prevent them, but they do help forecast threatened communities to secure property and take shelter.

IV. HOW CAN THEORIES BE MADE?

There is no agreed recipe for making theories.¹⁸ Some scholars use deduction, explanations from more general, already-established causal laws. Thus much

¹⁵For other examples see King, Keohane, and Verba, *Designing Social Inquiry*, p. 113, mentioning Talcott Parsons and David Easton's systems' analysis of macro-politics. On Easton see also Harry Eckstein, "Case Studies in Political Science," in Fred I. Greenstein and Nelson W. Polsky, ed., *Handbook of Political Science*, Vol. 7, Inquiry (Reading, Mass.: Addison-Wesley, 1975), pp. 79-137 at 90.

¹⁶Arguing the impossibility of a recipe is Hempel, *Philosophy of Natural Science*, pp. 10-18; also Milton Friedman in *Positive Economics* (Chicago: University of Chicago Press, 1953); constructing hypotheses "is a creative act of intuition, invention... the process must be discussed in psychological, not logical, categories; studied in autobiographical biographies, not treatises on scientific method, and promoted by maxim and example, not syllogism or theory. See also Shively, *Craft of Political Research*, pp. 163-166, noting the possibility of creating theories by induction and borrowing theories from other fields.

¹⁷Discussing this requirement of theory is Hempel, *Philosophy of Natural Science*, pp. 30-32.

¹⁸From there the theorist could move further by returning to deduction, e.g., deducing that conditions that inhibit war are a disadvantage for the offensive on the battlefield—are also causes of war.

theory is deduced from the assumption that people seek to maximize their personal economic utility. Others make theories inductively: they look for related specific phenomena, establish causation between them, and then 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.¹⁹

Several techniques can be used to aid inductive theory-making. One technique is to examine "outlier" cases, i.e., cases that are poorly explained by existing theories. Unknown causes must explain their outcomes. These causes can be identified by examining the case. Specifically, we select cases that lie furthest from the regression line expressing the relationship between the dependent variable and its known causes (i.e., "outlier" cases).²⁰ Candidate new causes will announce themselves as unusual characteristics of these cases, and as characteristics that are associated with the dependent variable within the case.²¹ We nominate these as candidate causes.²² We also cull the views of people who experienced the case or know it well and nominate their explanations as candidate causes.

The "method of difference" and "method of agreement" (proposed by John Stuart Mill) provide a second aid to inductive theory-making. In the method of difference the analyst compares cases with similar background characteristics and different values on the study variable (i.e., the variable whose causes or effects we seek to discover), looking for other differences between cases. These other cross-case differences are nominated as possible causes of the study variable (if we seek to discover its causes) or possible effects (if we seek its effects).²³

¹⁹From there the theorist could move further by returning to deduction, e.g., deducing that conditions that intensify competition for security--such as an advantage for the offensive on the battlefield--are also causes of war.

²⁰Another term for exploring outlier cases is "deviant case analysis." See Arund Liphart, "Comparative Politics and the Comparative Method," *American Political Science Review*, Vol. 65, No. 3 (September 1971), pp. 682-693 at 692.

²¹When making a theory, we select cases where the DV's causes are scarce yet the DV is abundantly present. This suggests that unknown causes are operating in the case, and that study of the case may reveal them.

²²When inferring a theory's antecedent condition (CV), we select cases where the DV's causes are abundant yet the DV is scarce or absent. This suggests that unknown antecedent conditions are absent in the case, and that study of the case may identify them.

²³For example, India, as a democracy with low levels of literacy, is an outlier from the regression line expressing the well-known relationship between democracy (the dependent variable) and levels of literacy (the independent variables.) Exploring the India case will uncover causes of democracy that operate independently of literacy and in addition to it.

²⁴I alter paragraphs in this section assume that the analyst seeks to fashion hypotheses on the causes of phenomena, neglecting the search for their effects. My advice can be adjusted to accommodate this goal, however, by substituting "independent variable" wherever "dependent variable" appears, and vice-versa, and by substituting "effects" for "causes."

Similar cases are picked to reduce the number of candidate causes or effects that more similar cases produce fewer candidates, making the real cause or effect spot.²⁴ Likewise, in the method of agreement the analyst explores cases with characteristics and similar values on the study variable, looking for other similarities between the cases, and nominating these similarities as possible causes or effects variable.²⁵

Third, we can select cases with extreme high or low values on the dependent and explore them looking for phenomena associated with the DV. If the DV is present abundance its cause should also be present in unusual abundance and should suggest against the case background. If the DV is absent its cause should also be prominent absence.

Fourth, we can select cases with extreme within-case variance on the dependent, and then explore the case looking for phenomena that co-vary with the DV on the DV vary sharply its cause should also vary sharply, standing out against static case background.

²⁴An example of using paired method-of-difference case studies for theory-making is Morris P. Fiorina, *Congress and the Washington Establishment* (New Haven: Yale University Press, 1977), chapter 4 (pp. 29-37). Fiorina explain why marginal congressional districts (i.e., "swing" districts where Democrats and Republicans congressional elections) were disappearing. To generate hypotheses he compared two districts highly similar but different in result: one district had always been and remained marginal, the other had changed from marginal during the 1960s. He nominated the key cross-district difference that he observed (greater constituency by the congressional incumbent in the newly non-marginal district) as a possible cause of the general decline. The growth of government, he theorized, had created opportunities for incumbents to win the voters' forming constituent service, and this bolstered incumbents who seized the opportunity.

²⁵I also had an early social science adventure inferring a hypothesis by method-of-difference case comparison. I was oblivious of J.S. Mill at the time). In 1969 I sought to explain why black political mobilization remained Deep South even after the passage of the 1965 Voting Rights Act. I inferred an explanation--holding that coercion by whites was retarding black mobilization--partly from Delphi-method interviews (see note 26, below) from a method-of-difference comparison.

I started by comparing two very similar black-majority Mississippi counties. Holmes and Humphries counties virtual twins on nearly all socio-economic dimensions except one: blacks had won county-wide elections in Holmes losing badly in next-door Humphries. This spurred my search for a second difference between them. It was Holmes had the Milleson project, a community of black handworkers who bought small farms through the New Security Administration in the 1940s. Humphries had nothing similar. As a result Holmes had far more blacks than Humphries. Further investigation (i.e., process tracing) revealed that these handworkers had played building Holmes County's black political organization. Interviews further suggested that fear of eviction among part farmers deterred their political participation throughout Mississippi, and the Milleson farmers were encouraged by their freedom from fear of eviction. A large-n test using all 29 black-majority Mississippi found a significant correlation between measures of black freedom from economic coercion and black political. This further corroborated the hypothesis that economic coercion depressed black political mobilization in Mississippi black belt, and suggested that such coercion might explain low levels of black mobilization in Deep South.

The results of this study are summarized in Lester M. Salamon and Stephen Van Evera, "Fear, Discrimination: A Test of Three Explanations of Political Participation," *American Political Science Review* 4 (December, 1973), pp. 1288-1306. Unfortunately, our article omits my Holmes county interview and pre-data. Still we behind the ears, I assumed that large-n tests were the only kind. It never occurred to me to go county as a case study.)

²⁶The method of difference is more efficient when the characteristics of available cases are quite homogeneous (most cases are similar). The method of agreement is preferred when the characteristics of cases are heterogeneous (most aspects of most cases are different).

Counterfactual analysis provides a fifth aid to inductive theorizing. The analyst examines history, attempting 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 motivating Nazi aggression: "how might the 1930s have unfolded if Hitler had died in 1932?" The greater the changes that one's analysis suggests would have followed from the changes posited, the more important one's analysis.

When analysts discover counterfactual analyses they find persuasive they have found theories they find persuasive, since all counterfactual analysis rests on theories: the analyst uses theories to predict how changed conditions would change events. If the analyst's colleagues 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 a general manner so that predictions can be inferred from it and tests can be performed. The analyst should ask: "what general causal laws are the dynamics I assert examples of?" The answer is a theory.

Counterfactual analysis merely helps us to recognize theories, not to make them. The theory that a counterfactual analysis uncovers exists in the theorist's subconscious before the theorist constructs the counterfactual. Otherwise the theorist could not construct it. Most people believe in more theories than they know. The hard part is to frame these theories explicitly, and to express them in general terms. Counterfactual analysis aids this process of framing and exposition.

Sixth, 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 ("communism 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.") These general theories can be tested. Such tests in turn can help resolve the policy debate. Theories inferred in this fashion are sure to have policy relevance and they merit close attention for this reason.

Seventh, the insights of actors or observers who experienced the event one seeks to explain are mined for hypotheses. Those who experience a case often observe important unrecorded data that is lost to later investigators. Hence they can suggest hypotheses that could not be inferred from direct observation alone.²⁶

²⁶ 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. However, interviews 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 mobilization. Further investigation unearthed substantial evidence to support their argument. (This hypothesis also emerged from a method-of-difference comparison of two Mississippi counties. See note 24.)

Finally, to mention an aid to deductive theorizing, theories can be fashioned by importing existing theories from one domain and adapting them to explain phenomena in another.²⁷ Thus students of misperception in international relations and students of political behavior have both borrowed theories from psychology. Students of political affairs have borrowed theories from the study of organizations. Students of international systems have borrowed theories (e.g., oligopoly theory) from economics.

V. HOW CAN THEORIES BE TESTED?

We have two basic ways to test theories: experimentation and observational tests come in two varieties: large-n and case study. Thus, overall universe of three basic testing methods: experimentation, observation using large-n, and observation using case study analysis.²⁸

A. "Experimentation." An investigator infers predictions from a theory; the investigator exposes one of two equivalent groups to a stimulus not exposing the other group. Are results congruent or incongruent with predictions? Congruence confirms the theory, incongruence infirms it.

B. "Observation." An investigator infers predictions from a theory; the investigator passively observes the data without imposing an external stimulus on the situation, and asks if observations are congruent with predictions. Predictions frame observations we expect to make if our theory is correct. They define expectations about the incidence, sequence, local structure of phenomena.³⁰ For instance, we can always predict that the structure of independent and dependent variables of valid theories should be as follows:

²⁷ Suggesting this technique is Shively, *Craft of Political Research*, p. 165.

²⁸ Deduction supplies a fourth way to evaluate theories. Using deduction to evaluate the hypothesis that 'a' causes 'b' would ask if 'a' and 'b' are examples of more general phenomena (A' and B') that are already known to cause 'b'. 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'. This is an assessment of theory see, e.g., Hempel, *Philosophy of Natural Science*, pp. 38-40 speaking of "theoretical" theories. A related discussion is ibid., pp. 51, speaking of "deductive-nomological" explanations and "covering" laws. The former are explanations that operate by deduction from general laws; the latter are general laws from which explanations are deduced.

Most "common sense" explanations are theories we believe because they are supported by deductions. However, a deductive evaluation is not a test of a theory. Rather, it applies a previously-tested law to a new case. Observational research designs are also called "quasi-experimental." See Donald T. Campbell and Julian H. Meehl, *Experimental and Quasi-Experimental Designs for Research* (Boston: Houghton Mifflin, 1963), p. 34.

³⁰ Use "prediction" to define expectations about the occurrence of phenomena in both the past and the future. 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'. This is an assessment of theory see, e.g., Hempel, *Philosophy of Natural Science*, pp. 7, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100. The former use "prediction" to refer to expectations about what the historical record will reveal, reserving "prediction" to refer to expectations about the future.

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

across time and space, other things being equal. Values on intervening variables that form the theory's explanation should also co-vary with the independent variable across time and space. Variance on the independent variable should precede in time related variance on the dependent variable. If a political theory is being tested, political actors should speak and act in a manner fitting the theory's logic (e.g., 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 (e.g. conflict simulations or psychology experiments). This leaves observation as our prime method of testing.

Two types of observational analysis can be performed:

1. **Large-N, or "statistical," analysis.**³¹ A large number of cases--usually several dozen or more--is assembled and explored to see if variables co-vary as the theory predicts.
2. **Case study analysis.** A small number of cases (as few as one) are explored in detail, to see if events unfold in the manner predicted and (if the subject involves human behavior) if actors speak and act as the theory predicts.³²

Which method is best? We should favor the method that allows the most strong tests (on strong tests see below, section VI, point 9; also Memo 2, below, section VII.) More tests is better than fewer; strong tests are better than weaker; many strong tests is best, as are methods that allow them. Most theories of international politics are best tested by case study methods because the structure of the international historical record, which serves as our data, usually lends itself better to deep study of a few cases than to exploration of many cases. As a result case studies often allow more and stronger tests than large-n methods. However, large-n analysis can be the better method if the data allow study of many cases. Experimentation is the least-valuable approach because experiments are seldom feasible.

VI. HELPFUL HINTS FOR TESTING THEORIES

Theory-testers should follow these injunctions:

1. Test as many of a theory's hypotheses as possible. Scholars sometimes hypotheses selectively, testing only those hypotheses that are borne out by evidence like the theory to begin with, or focusing on weak hypotheses if they dislike the theory. This is bad practice because it leaves the theory partly tested. A theory is tested by testing all its parts.

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

$$A \longrightarrow q \longrightarrow r \longrightarrow B$$

A complete test would evaluate the theory's prime hypothesis ($A \longrightarrow B$); its explanatory hypotheses ($A \longrightarrow q$, $q \longrightarrow r$, and $r \longrightarrow B$); and their hybrid combinations ($A \longrightarrow r$, $q \longrightarrow B$). Thus a three-link theory comprises a total of six testables. All should be explored if time and energy permit.

2. Test as many predictions of each hypothesis as possible. This requires infer as many predictions as possible from your hypothesis. Consider what your hypothesis predicts across time and space (i.e., across regions, groups, individuals). Consider also what general decision process (if any) it predicts specific individual speech and action it predicts. Most hypotheses make several predictions, so don't be quickly content to rest with one.

Predictions frame observations you expect to make if the theory is valid. The expectations about the incidence, sequence, location, and structure of phenomena framing tautological predictions that forecast simply that we expect to observe in operation ("if the theory is valid, I predict we will observe its cause causing"). Thus the hypothesis that "democracy causes peace" yields the following tautological prediction: "we should observe democracy causing peace." A non-tautological prediction would be: "we should observe that democratic states are involved in fewer authoritarian states."

3. Explain and defend the predictions you infer. Scientific controversies from disputes over which predictions can be fairly inferred from a theory and not be. We then see scientists agree on the data but dispute what it means be dispute what the tested theories predict. Theorists can minimize such disputes explaining and defending their predictions.

Predictions can be either general (broad patterns are predicted) or specific facts or other single observations are predicted). General predictions are inferred and are used to test, general hypotheses ("if windows of opportunity and windows drive states to war, states in relative decline should launch more than their share). Specific predictions are inferred from, and are used to test, general hypotheses windows of opportunity and vulnerability drive states to war, we should see Japan 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

³¹Printers on large-n analysis include Babbie, *Practice of Social Research*; Shively, *Craft of Political Research*; William G. Cochrane, *Planning and Analysis of Observational Studies* (New York: Wiley, 1983); 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); David Freedman et al., *Statistics*, 2nd ed (New York: Norton, 1991).

³²Landmark writings on the case study method are listed in Memo 2, "What Are Case Studies? How Should They Be Performed?", in this working paper, footnote 1.

find records of Japanese decision makers citing a closing window as reason for war³¹), and non-generalized specific explanations ("If Japan started the Pacific War, we should find photographs of Japanese planes flying over Pearl Harbor on December 7, 1941").³²

4. Select data that represents, as accurately as possible, the domain of the test. When using large-n test methods, select data that represents the universe defined by tested hypotheses. When using case study methods, select data that represents conditions in the cases studied. Even data that represents the domain of the test only crudely can still be useful.³⁴ However, the more accurate the representation, the better. Choosing evidence selectively--e.g., favoring evidence that supports your hypothesis over disconfirming counter-evidence--is disallowed, since it violates the principle of accurate representation.

This rule is almost a platitude, but older international relations literature 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 case study.

5. Consider and evaluate the possibility that an observed relationship between two variables is not causal, and results instead from the effect of a third variable.³⁵ Two variables may co-vary because one causes the other, or because a third variable causes both. For example, monthly sales of ice cream and baseballs correlate closely in the northern U.S. but neither causes the other. Instead, warm weather causes both. Controls on the effects of such third variables should be considered or introduced before concluding that correlation between variables indicates causation between them.

6. When interpreting results:

- a. If you flunk (or pass) a theory, do not assume *a priori* that the same verdict applies to all similar theories. Each theory in a theory family (e.g., the neoclassical economic theory family, Marxist theories of imperialism, Realist theories of international relations, etc.) 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 variants of the same more general theory, and your test has refuted or corroborated that general theory.
- b. If you flunk (or pass) one hypothesis in a multi-hypothesis theory, remember that this says nothing about the validity of other hypotheses in the theory. Each hypothesis must be tested on its own before we can assess its validity.
- c. Don't reject flunked theories before considering if they can be repaired. Flunked theories often contain valid hypotheses. They can be salvaged and incorporated into a new theory.

³¹On the difference between hypotheses, generalized specific explanations, and non-generalized specific explanations see section I above and section VII below.

³⁴Arguing for and illustrating the utility of "rule of thumb" tests using data that only crudely represent a larger domain is John J. Mearsheimer, "Assessing the Conventional Balance: The 3:1 Rule and Its Critics," *International Security*, Vol. 13, No. 4 (Spring 1989), pp. 54-89 at 56-62.

³⁵A discussion is Babbie, *Practice of Social Research*, pp. 396-409.

7. Theories are repaired by replacing disconfirmed hypotheses with new explanatory hypotheses proposing a different intervening causal process, or by narrowing the theory's claims. We narrow a theory's claims by adding new antecedent conditions (i.e., condition variables, or CVs), so the theory no longer claims to govern the cases it was flunked. This allows the flunked test to be set aside. The theory is more modest but passes its tests.

8. Theories can be tested against the null hypothesis ("does this theory explainatory power?" or against each other ("does this theory have more or less explanatory power than competing theories?").³⁶ Both test formats are useful and legitimate should not be confused. Theories that pass all their tests against the null should be named the leading theory without further investigation; they can still flunk tests competing theories. Conversely, theories that lose contests against competitors should be dismissed altogether. They still may have some explanatory power, and their explanatory power are valuable even if other theories have more.

9. Strong tests are better than weak tests, and the results of strong tests carry more weight than the results of weak tests.³⁷

A strong test is one whose outcome is unlikely to result from any factors other than operation or failure of the theory. Strong tests evaluate predictions that are *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 events that must inexorably occur if the theory is valid. If the prediction fails the test, since failure can arise only from the theory's non-operation. A *unique* prediction is one that is 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 alternative except the theory's action. If the prediction succeeds the theory is strongly corroborated because competing 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 certainty or uniqueness falls, the strength of the test falls. Tests of predictions that have little certainty or uniqueness are weakest, and are the least predictive.

³⁶Imre Lakatos likewise distinguishes "a two-cornered fight between theory and experiment" and "three-cornered fight between rival theories and experiment." His "two-cornered fights" are tests against the null, his "three-cornered fights" include a test against the null and a theory-against-theory test. Imre Lakatos, "Falsification and the Method of Scientific Research Programmes," in Imre Lakatos and Alan Musgrave, eds., *Criticism and the Growth of Mathematics* (Cambridge: Cambridge University Press, 1970), pp. 91-196 at 115. Works formatted as two-cornered fights include many studies on democratic peace theory, e.g., Steve Chan, "Mirror, Mirror on the Wall ... Are the Freer Countries More Peaceful?" *Journal of Conflict Resolution*, Vol. 28, No. 4 (December 1984), pp. 617-648; Erich Weede, "Do Democracies Fight?" *Journal of Conflict Resolution*, Vol. 28, No. 4 (December 1984), pp. 649-664. A study of three-cornered fights is Barry R. Posen, *The Sources of Military Doctrine: Britain, France, and Germany in the Great War* (Ithaca, N.Y.: Cornell University Press, 1984). For more on the topic see Hempel, *Philosophy of Science*, pp. 25-28 (discussing "crucial tests").

³⁷Discussions of strong tests include Eckstein, "Case Study and Theory," pp. 113-131, discussing what he calls "strong case studies" (his term for cases supplying strong tests); and Arthur L. Stinchcombe, *Constructing Social Worlds* (New York: Harcourt, Brace & World, 1968), pp. 20-22.

Strong tests are preferred, but tests can also be hyper-strong, i.e., unfair to the theory. These tests should be avoided or interpreted carefully. For example, one can subject a theory to tests where countervailing forces are present that counteract its predicted action. Passage of such tests is impressive because it shows the theory's cause has high impact. However, such tests are unfair to the theory unless the investigator gives the theory bonus points for the extra hardship it faces. Otherwise a valid theory may flunk its tests even as it operates because countervailing factors mask its action. Such a test misleads by recording a false negative.

Another form of hyper-strong test evaluates theories under circumstances where they are likely to fail because antecedent conditions required for their success are missing. 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. However, such tests are unfair as a measure of a theory's basic validity, since they assess it against claims that it does not make.³⁸

10. A theory is tested by asking if the theory's predictions are confirmed by empirical evidence, not by asking how many cases the theory can explain. A theory may explain few cases because its independent variable 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, *ceteris paribus*. However, even theories that explain very few cases are valuable if these cases are important and the theory explains them well.

11. A theory is not tested by assessing the validity of its assumptions (i.e., 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 merely because they flunk such tests.

However, the validity of a theory's assumptions does affect its utility. Assumptions that never hold give rise to theories that operate only in an imaginary world, and thus cannot explain reality or generate policy prescriptions.³⁹ The most useful theories are those whose assumptions match reality in at least some important cases.

³⁸ Advocates of testing theories against "least-likely" cases—i.e., 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 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.

³⁹ 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 the nature of the theory's inner workings is uninteresting and knowledge about these inner workings is correspondingly unuseful, but this is seldom the case in the study of politics.

VII. HOW CAN SPECIFIC EVENTS BE EXPLAINED?

Ideas framing cause and effect come in two broad types: theories and specific explanations. Theories are cast in abstract terms and could apply to more than one case ("expansionism causes war"). Specific explanations explain discrete events—"wars, interventions, empires, revolutions, or other single occurrences (e.g., expansionism caused World War II," or "an asteroid impact caused the extinction of dinosaurs.") The framing and testing of theories is covered above, but how should we test specific explanations?⁴⁰ Four questions should be asked:

1. Does the explanation exemplify a valid general theory (i.e., a covering law) that assesses the hypothesis that 'a' caused 'b' in a specific instance we first assess the hypothesis' general form ("A causes B"). If 'A' does not cause 'B', we can rule out all instances of specific instances of 'B' that assert that examples of 'A' were the cause, the hypothesis that 'a' caused 'b' in this case.

The argument that "masturbation caused Caligula's madness" is assessed whether, in general, masturbation causes madness. If the hypothesis that "masturbation causes madness" has been tested and flunked, we can infer that masturbation cannot cause Caligula's madness. The explanation fails because covering law is false.

Generalized specific explanations are preferred to non-generalized specific explanations because we can measure the conformity of the former but not the latter covering laws. Non-generalized specific explanations must be re-cast as general 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 covering theory on which the explanation rests is greater than zero. Even if 'A' is a cause of 'B', it cannot explain instances of 'B' that occur when 'A' is absent.

⁴⁰ The role of theories in historical explanation has long been debated by historians and philosophers of social science. See Hempel, "The Function of General Laws in History," the landmark work in the debate, and other reactions see Martin and McIntyre, *Readings in the Philosophy of Social Science*, pp. 55-156. A more recent work is Clayton Roberts, *The Logic of Historical Explanation* (University Park, Pa.: Pennsylvania State University Press, 1996). See also Eckstein, "Case Study and Theory," pp. 99-104, discussing "disciplined-configurative" meaning case studies that aim to explain the case by use of general theories.

⁴¹ 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.

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 pre-capitalist empires. Asteroid impacts may cause extinctions, but cannot explain extinctions that occurred in the absence of an impact.⁴²

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 non-rabid dog cannot explain a rabies case.

4. Are the covering law's intervening variables observed in the case? Variables that link the covering law's posited cause and effect should be evident and appear in the proper order.⁴³ Actors should be observed responding to the independent variable or to stimuli that it produces. To assess the hypothesis that 'a' caused 'b' in a specific instance we first verify the relevant covering law--'A' causes 'B'--and then ask how 'A' causes 'B'. If 'A' causes 'Q' and 'b' should appear in sequence after 'a' appears. We should further find documentary or other evidence that actors who cause 'q' do so in response to the appearance of 'a', and those causing 'b' are responding to 'q'. We should find signs of other phenomena that 'A' and 'Q' tend to cause. If the Russian mobilization caused World War I, it should precede other triggering acts or events (e.g., the German invasion of Belgium), and those responsible for these triggering acts should explain in private records that they responded to Russia's mobilization. Etc.

This fourth step is necessary because the first three steps are not definitive. If we omit step four, 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.⁴⁴

⁴²Thus if an asteroid killed the dinosaurs 65 million years ago we should find evidence of a very large asteroid impact at that time (e.g., remnants of a large impact crater of the right age and size; iridium, soot, and shocked quartz, which deduction suggests a large impact would produce in quantity, in 65-million-year-old sediment layers around the world, etc.) Observation of this evidence would corroborate the impact theory, while failure to observe it would inform the theory. (In fact this prediction is fulfilled. Quantities of iridium, soot, and shocked quartz are found just where the impact theory predicts. The theory suffered for years because the predicted impact crater could not be found, but the discovery of the Chicxulub crater in Mexico has provided a likely candidate.)

The debate over the dinosaur extinction nicely illustrates the inference and framing of clear predictions from specific explanations. On that debate see Walter Alvarez and Frank Asaro, "An Extraterrestrial Impact," *Scientific American* (October 1990), pp. 78-84; Vincent Courtillot, "A Volcanic Eruption," in *ibid.*; and William J. Broad, "New Theory Would Reconcile Views on Dinosaurs' Demise," *New York Times*, December 27, 1994, p. C1.

⁴³Thus 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 theorize that an impact would kill by spraying the globe with molten rock, triggering global forest fires that blacken the skies with smoke, shutting out sunlight and freezing the earth. If so, the soot these fires would generate should be found (and in fact is found) in 65-million year old sediment worldwide. Others theorize that an impact would kill by triggering vast volcanic eruptions at a point directly opposite the impact on the globe; these eruptions would cause catastrophic climatic change. (Evidence of such eruptions is found in India's Deccan Traps.)

⁴⁴The cause of probabilism in probabilistic causal laws usually lies in the occasional absence of required antecedent conditions that we have not yet identified.

We also need to test the explanation's within-case predictions as a hedge against billity that our faith in the covering law is misplaced, and that the "law" is in fact these two reasons, the better the details of the case conform to the detailed within-dictions of the explanation the stronger the inference that the explanation explains

Analysts are allowed to infer the covering law that underlies the specific event of a given event from the event itself. The details of the event suggest a specific theory; that explanation is then framed in general terms that allow tests against a database; these tests are passed; and the theory is then re-applied to the specific case. general theory-testing and specific case-explaining can be done together and each other.

VIII. METHODOLOGY MYTHS

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

1. "Evidence infirming theories transcends in importance evidence confirming them." Karl Popper and other falsificationists argue that "theories are not verifiably true, but only falsifiable,"⁴⁵ and that tests infirming a theory are far more significant than tests confirming it.⁴⁷ Their first claim is narrowly correct, their second is not. Theories are proven absolutely because we cannot imagine and test every prediction they make. The possibility always remains that an unimagined prediction will fail. By infirming tests can decisively refute a theory. It does not follow that infirming tests can decisively confirm a theory. Strong confirming tests can give us high confidence in a theory. If that theory later flunks a test of a previously untested prediction, by means that the theory requires previously unidentified antecedent conditions. We react by reframing the theory to include the antecedent condition, thus infirming the scope of the theory's claims to exclude the flunked test. In Popper's terminology, we have a new theory. However, all the tests passed by the old also confirm the new theory in very strong shape at birth. Thus confirming tests tell us a great deal--about theory, about its repaired replacement, and about any later re-repaired versions. Contrary argument stems partly from his strange assumption that once theories are promptly accepted,⁴⁸ hence evidence in their favor is unimportant merely reinforces a preexisting belief in the theory. The opposite is more often true: new ideas face hostile prejudice even after confirming evidence accumulates.

⁴⁵Less convinced of the need for this last step is Hempel, "Falsification of General Laws," who rests with the first and omits the fourth. Hempel assumes that his covering laws are deterministic (not probabilistic) and are not subject to the same social science laws as probabilistic and most are poorly established. If so, deducing the veridicality of a covering law from the first three steps alone is unreliable, and we should also seek empirical verification of the covering law's causal process in fact occurred before reaching final conclusions.

⁴⁶Karl R. Popper, *The Logic of Scientific Discovery* (London: Routledge, 1959), p. 252. A criticism of Popper's position is King, Keohane, and Verba, *Constructing Social Inquiry*, pp. 100-103.

⁴⁷In a friendly summary of falsificationism David Miller writes that falsificationists "the passing of tests is a job of difference to the status of any hypothesis, though the failing of just one test may make a great deal of difference to the status of any hypothesis." Popper's Solution of the Problem of Induction," in Paul Levinson, *Truth* (Atlantic Highlands, N.J.: Humanities Press, 1988), p. 22.

⁴⁸See King, Keohane, and Verba, *Designing Social Inquiry*, p. 100.

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."⁴⁹ This claim is too sweeping. It applies only to theories that fail some tests but retain some explanatory power. These theories should be retained 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.⁵⁰ Many science programs--e.g., medical research--advance by routinely testing theories against null hypotheses and rejecting those that fail whether or not replacements are ready.

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 his or her own explanation could not meet. This suggests in turn that the standard was too high, e.g., 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's wrong before knowing what's right.

3. "The evidence that inspired a theory should not be re-used to test it." This argument⁵¹ 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.⁵² 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 hardly a fixed rule.⁵³ Its purpose is to prevent scholars from choosing confirming tests while omitting infirming ones.

⁴⁹Lakatos, "Falsification and the Methodology of Scientific Research Programmes," pp. 119, 122.

⁵⁰An early reader of this memo suggests that Lakatos means to argue 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' arguments are well-hidden in tortured prose that gives new meaning to the phrase "badly written," and no reading of such abominable writing is ever certain or final.

⁵¹Raising 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*, Vol. 2 (Greenwich, Conn.: JAI Press, 1985), pp. 21-38 at 38; David Collier, "The Comparative Method," in Ada W. Funder, ed., *Political Science: The State of the Discipline*, 2nd ed. (Washington, D.C.: American Political Science Association, 1993), pp. 105-120 at 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).

⁵²A discussion is Hempel, *Philosophy of Natural Science*, pp. 37-38.

⁵³*Mea culpa*: I broke this rule myself. I inferred the core of my version of offense-defense theory from study of the Arab-Israeli conflict, but I did infer major elements from study of the outbreak of World War I. Then I later offered the 1914 case as evidence in favor of the theory. See Van Evera, *Cases of War*, vol. 1, forthcoming from Cornell University Press, chapter 7. My thought process was iterated and reciprocal: investigation of 1914 produced hypotheses that more investigation of 1914 corroborated. Hence the overlap of 1914 test data and 1914 data used for theory-inferencing was not complete, or perhaps even very high. Nevertheless, had it been high I think no harm would have been done as long as the test was designed and reported with integrity. Moreover, removing 1914 from the menu of test cases would have removed what is, for several reasons, the best test case, causing a loss of important evidence.

But imposing blind-test rules on social science is in fact impossible because investors usually know something about their data before they test their theories; hence they have a good idea what tests will show even if they exclude data that inspired their theory. Hence we need other barriers against test-fudging.⁵⁴ 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 that you seek to explain (e.g., wars) without also choosing cases of the contrary (p) students of the case method often repeat this warning.⁵⁵ It is false. Selection of dependent variable is appropriate under three common conditions:

- If conditions in selected cases can be compared to a known average situation. The average situation is often sufficiently well-known not to require further detective study. If so, we can compare cases selected on the dependent variable to known normal conditions. There is no need for full-dress case studies to pinpoint sharper points of comparison.⁵⁷
- If the cases have large within-case variance on the study variable, permitting congruence procedures.
- If cases are sufficiently data-rich to permit process tracing.⁵⁸

Neither congruence procedure nor process tracing requires comparison to specific external cases, hence failure to select diverse points of comparison does not create problems.

⁵⁴Moreover, a blind-test requirement would generate a preposterous double-standard in the right to use evidence: data would be forbidden as test material to some scholars (because they inferred the theory from it) while being open to others. How would this rule be administered? Who would record which scholars had used which data for their theory, and hence were barred from re-using it for testing? Would we establish a central registry of hypotheses whose risks would record the origins of their ideas? How would we verify and penalize failure to accurately record by theory with this registry? How would we deal with the many scholars who are not really sure where their hypotheses come from?

⁵⁵See Barbara Geddes, "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Political Analysis," Vol. 2 (1990), pp. 131-150; also King, Keohane, and Verba, *Designing Social Inquiry*, pp. 108-132, 137-138, 140-149. King et al. warn that "we will not learn anything about causal effects" from studies of cases with no variation on the dependent variable; they declare that the need for such variation "seems so obvious" that it hardly needs to be mentioned, and they conclude that research designs that lack such variation "deal with a world that is not the world we live in" (pp. 129-130). A criticism is Ronald Rogowski, "The Role of Scientific Theory and in Social-Scientific Inference," *American Political Science Review*, Vol. 89, No. 2 (June 1995), pp. 467-470. He notes that King, Keohane and Verba's strictures point to a "needlessly inefficient path of social-scientific inquiry," and distance to these strictures "may paralyze, rather than stimulate, scientific inquiry" (p. 470). On Geddes and King, and Verba see also David Collier and James Mahoney, "Insights and Pitfalls: Keeping Selection Bias in Perspective," U.C. Berkeley, February 6, 1996.

⁵⁶Thus Lijphart notes the "implicitly comparative" nature of some single case studies. "Comparative Politics: A Comparative Method," pp. 692-693.

⁵⁷Thus the erring scholars that Geddes identifies erred because they misconstrued the normal worldwide background of the key independent variables, e.g., intensity of labor repression, that they studied.

⁵⁸On congruence procedure and process tracing see Memo 2, "What Are Case Studies? How Should They Be Performed in this Working Paper," sections II B and II C.

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.⁵⁹ This criteria 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 sought object lies in the 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.⁶⁰ 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 answers will be hard to find.⁶¹ 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 the theory-testing." James Fearon suggests this argument.⁶² 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 above, section V). But counterfactual statements are not data and cannot substitute for empirical data in theory-testing.

MEMO 2: WHAT ARE CASE STUDIES? HOW SHOULD THEY BE PERFORMED?

A large literature on the case study method has appeared in recent years,¹ but the literature remains spotty. No complete catalogue of research designs for case studies has emerged.² No textbook covers the full gamut of study design considerations.³ The *Encyclopedia of Case Studies* is a soup-to-nuts cookbook on the case method for beginning practitioners, and most social science methodology texts slight or omit the case study method.⁴ According to the *Handbook of Case Study*, "The case study method is a research strategy that involves the study of a single individual, group, or organization in depth. It is a method of inquiry that allows the researcher to explore the complexities of a particular case in a way that is not possible through other methods." According to the *Handbook of Case Study*, "The case study method is a research strategy that involves the study of a single individual, group, or organization in depth. It is a method of inquiry that allows the researcher to explore the complexities of a particular case in a way that is not possible through other methods." I focus on assessing the case study method and offering practitioners do-it advice for beginners doing case studies.

1. CASE STUDIES IN PERSPECTIVE

As noted above in Memo 1, we have two basic ways to test theories: experiments and observation.⁵ Observational tests come in two varieties: large-n and case study. Overall we have a universe of three basic testing methods: experimentation, observation using large-n analysis, and observation using case study analysis.

Which testing method is best? Is case study inferior to other methods?

¹A good survey of the case study literature is David Collier, "The Comparative Method," in Ada W. Finifter, ed., *The State of the Discipline*, 2nd ed. (Washington, D.C.: American Political Science Association, 1987), pp. 120-121. Landmark writings on the case method include Alexander L. George and Timothy J. McKeown, "Case Studies of Organizational Decision Making," in *Advances in Information Processing and Organization* (Greenwich, Conn.: JAI Press, 1985), pp. 21-58; Arvid Liphart, "Comparative Politics and the Case Study Method," in *American Political Science Review*, Vol. 65, No. 3 (September 1971), pp. 682-693; Harry Eckstein, "Case Studies of Inquiry (Reading, Mass.: Addison-Wesley, 1975), pp. 79-137; Robert K. Yin, *Case Study Research: Design and Methods*, 2nd ed. (Thousand Oaks, CA: Sage, 1994). A more developed but less accessible discussion is Alexander L. George, "Case Studies and Theory Development," paper presented to the Second Annual Symposium on Information Processing in Organizations, Carnegie-Mellon University, Pittsburgh, Pa., October 15-16, 1990. A statement is Alexander L. George, "Case Studies and Theory Development: The Method of Structured Comparison," in Paul Gordon Lauren, ed., *Diplomacy: New Approaches in History, Theory, and Policy* (New York: St. Martin's Press, 1979), pp. 43-68.

²Yin, *Case Study Research*, pp. 18-19.

³Yin, *Case Study Research*, p. 18. Useful steps toward such a text are *ibid.*, and Gary King, Robert O. Sidney Verba, and Kenneth J. Bratton, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press, 1994).

⁴Yin, *Case Study Research*, pp. 13, 18-19; Jennifer Platt, "Case Study" in *American Methodological Theoretical Sociology*, Vol. 40, No. 1 (Spring 1992), pp. 17-48 at 42-43. Illustrating are, e.g., Earl Babbie, *The Practice of Social Research*, 7th ed. (Belmont, Calif.: Wadsworth, 1995); Julian L. Simon and Paul Burslein, *Basic Research in Social Science*, 3rd ed. (New York: Random House, 1985); Kenneth D. Bailey, *Methods of Social Research* (New York: Free Press, 1994); David Doolley, *Social Research Methods*, 3rd ed. (Upper Saddle River, N.J.: Prentice-Hall, 1989); Norman K. Denzin, *The Research Act*, 3rd ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1989). Babbie mentions once (p. 280); Simon and Burslein give case study methods a 2-page mention (pp. 37-38); Bailey has a 1-page mention (pp. 301-303); Doolley has a chapter on "qualitative research" (pp. 257-274) but no direct mention of case studies.

⁵See Memo 1, "Hypotheses, Laws and Theories: A User's Guide," this working paper, section V.

⁵⁹King, 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.

⁶⁰See, e.g., the last several decades of *The American Political Science Review*.

⁶¹Moreover, tests that are difficult for the time being may become feasible as new tests are devised or new data emerges. This is another reason to keep hard questions on the agenda. Thus theories of the Kermelin's conduct under Stalin were hard to test before the Soviet collapse but later became more testable.

⁶²James D. Fearon, "Counterfactuals and Hypothesis Testing in Political Science," *World Politics*, Vol. 43, No. 2 (January 1991), pp. 169-195 at 171 and *passim*.

Case studies have long been thought the weakest of these three testing methods, for two reasons.⁶ First, some argue that case studies provide the least opportunity to control for the effect of perturbing third variables. In this view experiments are the best method (the investigator eliminates the possible effect of omitted variables by exposing the group to only one stimulus, while holding the others constant). Large-n analysis is next-best, because the investigator can run partial correlations to control the effect of specific omitted variables, and can rely on the randomizing effect of examining many cases to reduce the effects of other omitted variables. Studies of one or a few cases are worst, because the data is unrandomized and partial correlations are infeasible--unless many cases are examined, in which case we are back to large-n analysis.⁷

This criticism of case studies is unfair, however. Case studies offer two fairly strong methods for controlling the impact of omitted variables. First, tests of within-case predictions, using a multiple "congruence procedure"⁸ or a "process tracing" methodology,⁹ can achieve strong controls by conducting the test against uniform background conditions.¹⁰ Most cases offer a backdrop of fairly uniform case conditions, and many cases allow a number of observations of values on the independent and dependent variables. If case conditions are uniform we can discount third-variable influence as a cause of observed within-case co-variance between values on IV and DV. (The uniform background conditions of the case create a semi-controlled environment that limits the effects of third variables by holding them constant.) If observations are numerous we can further discount omitted

⁶See e.g., Yin, *Case Study Research*, pp. 9-11, who notes the "traditional prejudice against the case study strategy" and the "disdain for the [case study] strategy" held by many researchers (p. 9). As Yin further notes, social science methodology texts reflect this disdain by neglecting or omitting the case method: "most social science textbooks have failed to consider the case method a formal research strategy" at all" (p. 13). Randy Stoecker likewise notes the disrepute of case studies among sociologists, who "see the case study as barely better than journalism." Randy Stoecker, "Evaluating and Rethinking the Case Study," *The Sociological Review*, Vol. 39, No. 1 (February, 1991), pp. 88-112 at 88. See also Jacques Hamel with Stephanie Dufour and Dominic Fortin, *Case Study Methods* (Newbury Park, CA: Sage, 1993), pp. 18-28.

⁷Advancing this view is Lipihart, "Comparative Politics and the Comparative Method," pp. 683-684; similar is Neil J. Smelser, "The Methodology of Comparative Analysis," in Donald P. Warwick and Samuel Osterson, eds., *Comparative Research Methods* (Englewood Cliffs, N.J.: Prentice Hall, 1973), pp. 45, 57. A synopsis of Lipihart is Collier, "Comparative Method," p. 106-108. See also Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Boston: Houghton Mifflin, 1963), p. 6, who claim that single case studies are "of almost no scientific value." But see further Campbell's later retraction: Donald T. Campbell, "Degrees of Freedom and the Case Study," in Donald T. Campbell, *Methodology and Epistemology for Social Science: Selected Papers* (Chicago: University of Chicago Press, 1988, first pub. 1974), pp. 377-388; and noting this retraction, Collier, "Comparative Method," p. 115.

⁸In a multiple congruence procedure the investigator explores the case looking for congruence or incongruence between observed and predicted values on several or more measures of the independent and dependent variables of the test hypothesis. See this memo, section II B below. To test a theory fully one would look for congruence or incongruence between values on independent and dependent variables, between independent and intervening variables, between intervening variables (if there are several), and between intervening and dependent variables.

⁹On "process tracing" see George and McKeown, "Case Studies and Theories," pp. 34-41; George, "Case Studies and Theory Development," pp. 18-19; and this memo, section II C, below. George and McKeown use "process tracing" to refer to a tracing of "the decision process by which various initial conditions are translated into outcomes." "Case Studies and Theories," p. 35. I use the term more broadly, to refer to the tracing of any causal process by which initial conditions are translated into outcomes. Thus my definition includes the tracing of both decision processes and also causal processes that do not involve decisions. We might reserve "decision-process tracing" to capture George and McKeown's more narrow meaning.

¹⁰Noting the controls that congruence procedure and process tracing allow (and referring to them jointly as "pattern matching") is Campbell, "Degrees of Freedom and the Case Study," p. 380.

variables or measurement error as causes, and infer that co-variance between IV signals causation.¹¹ Second, the effects of omitted variables can also be controlled selecting for study cases with extreme (high or low) values on the study variable. The number of third factors with the strength to produce the result that the test predicts, which lowers the possibility that omitted variables account for passed test

A second criticism of case studies has more merit, but applies only to single case studies. Critics note that a single case study is a poor laboratory for identifying antecedent conditions, i.e., background conditions that activate or magnify its activity noted above, most cases provide a backdrop of fairly uniform case conditions. Uniformity masks the importance of antecedent conditions that the theory requires, antecedent condition does not vary, hence it causes no telltale variance on the DV theory that passes a single case study test with flying colors may require rare conditions, and hence have little explanatory range,¹⁴ but this weakness can remain from an investigator that studies only one or two cases. (Thus a strength of method is also a weakness. The uniformity of case background conditions conceals effects of third variables but also masks antecedent conditions.) The identity and variance of antecedent conditions emerges more clearly from large-n studies. They cases that lack these conditions, which emerge as outliers that exhibit the theory without its predicted outcome. The existence of outliers signals that the theory needs a single case offers no parallel method for uncovering antecedent conditions. If these antecedent conditions can be uncovered by doing more case studies, so this value in the case method is repairable.¹⁵

The case method has two strengths that offset this weakness. First, tests performed with case studies are often strong, because the predictions tested are quite uniform. These predictions are not made by other known theories.¹⁶ Specifically, case studies test the test of predictions about the private speech and writings of policy actors. Off-

¹¹This logic applies to analysis of any hypothesized causal relationship, e.g., between IV and DV, DV and DV, as well as IV and DV.

¹²On this technique see below, this memo, section II B. A third means of omitted-variable control in case studies is the method of controlled comparison, using John Stuart Mill's "method of difference," but this is a fairly weak section II A, this memo, below.

¹³The process of defining and measuring the prevalence of the antecedent conditions is often referred to as theory's "external validity," meaning tests "establishing the domain to which a theory can be generalized." *Case Study Research*, p. 33. Tests for external validity contrast with tests of "internal validity," which address the theory to explain a given case. See, e.g., Yin, *Case Study Research*, pp. 33, 35-36; Collier, "Comparative Method," p. 113. I avoid these binary categories because they omit an important third type of validity--directly testing the theory to pass tests in a given case.

¹⁴On explanatory range see Memo 1, "Hypotheses, Laws and Theories: A User's Guide," in this working paper, section II.

¹⁵Methods of inferring and testing antecedent conditions with case studies are discussed in this memo, section II B below. Criteria for selecting these additional cases is discussed in sections IX 2c-2d IX 3c-3d IX 6c-6d IX 7a-7b.

¹⁶A test is strong if it evaluates a unique prediction (i.e., a forecast not made by other known theories) because a theory's fulfillment cannot be explained except by the theory's action. Tests are also strong if they evaluate conditions (i.e., forecasts that are unequivocal and must occur if the theory is valid.) The strongest tests evaluate conditions that are both unique and certain. On strong and weak tests see this memo, below, section VII, and Memo 1, "Laws and Theories: A User's Guide," in this working paper, section VI 9.

predictions are singular to the theory that makes them; no other theory predicts the same thoughts or statements. Case studies are the best format for capturing such evidence. Hence case studies can supply quite decisive evidence for or against political theories. Often this evidence is more decisive than large-n evidence.

Second, inferring and testing explanations that define how the independent causes the dependent variable is often easier with case study than large-n methods. If case study evidence supports a hypothesis, the case can then be further explored to deduce and test explanations detailing the hypothesis' operation. Most important, one can "process trace," i.e., examine the process whereby initial case conditions are translated into case outcomes. How does the theory work? Tracing process can tell us. Congruence procedures can also illuminate explanations. (More on process tracing and congruence procedures below.) Both procedures are fairly easy to perform after a case has been initially studied because the background spade-work on the case--defining the case parameters, establishing the chronology, etc.--has already been done. In contrast, a large-n test of a hypothesis provides little or no new insight into the causal process that comprises the hypothesis' explanation. It does not generate data from which explanations might be inferred, and the investigator must do large new work (measuring intervening variables) to test explanatory hypotheses. Overall, large-n methods tell us more about whether hypotheses hold than why they hold.

Thus the case method is a strong method of testing theories, and a weaker means of identifying antecedent conditions that animate theories. Is a theory valid? How does it operate? Case studies can give clear answers. How broad is the range of cases that the theory governs? Case studies say little unless several are performed.

Which method of inquiry--experiment, large-n or case study--is superior? The answer depends on circumstances. The best method is the one that allows the most strong tests of theories and best reveals their antecedent conditions. This depends, in turn, on the structure of the data in the domain being studied. In the study of international relations case studies are usually more useful than large-n approaches because the structure of the international historical record, which serves as our data, better lends itself to deep study of a few cases than to exploration of many cases.¹⁷ This is because we often have extensive records of a few cases, but thin records of the rest. As a result, detailed study of a few cases can allow quite reliable tests of many predictions, while large-n methods do not allow reliable tests because reliable large-n databases are hard to assemble due to gaps in the historical record on more obscure cases. The marked shortcoming of single case studies--their inability to identify antecedent conditions--can be addressed by performing several studies. However, large-n analysis can be the better method if the data allow study of many cases while being poorly structured for in-depth study of these cases. Large-n analysis is also useful for identifying antecedent conditions that single case studies have failed to identify. (Experimentation is the least-valuable approach because experiments are seldom feasible.)

II. TESTING THEORIES WITH CASE STUDIES¹⁸

Case studies offer three formats for testing theories: controlled comparison, process procedure, and process tracing. Controlled comparison uses across-cases comparative observations to test theories. Congruence procedures are of two types, with comparative observations to test theories. The other using within-case observations. Process tracing tests theories using within-case observations. Comparison procedure and process tracing are stronger test methods than controlled comparison, three are also used to create theories and to infer and test antecedent conditions, below.)

In each format the same three steps are followed: (1) state the theory; (2) state observations about what we should observe in the case if the theory is valid, and what we observe if it is false; (3) explore the case (or cases) looking for congruence or incongruence between expectation and observation.

A. Controlled Comparison.¹⁹ The investigator explores paired observations two or more cases, asking if values on the pairs are congruent or incongruent with the tested theory's predictions. For example, if values on the independent variable are higher in case 'A' than case 'B', values on the dependent variable should also be higher in case 'A' than 'B'.

Case selection follows John Stuart Mill's "method of difference" or of agreement.²⁰ In the method of difference the investigator chooses cases with similar general characteristics and different values on either the dependent or independent variable, then asks if values on the other variable vary accordingly in corresponding fashion. Similar cases are picked to control for the third variables: the more similar the cases, the less likely that the action variables explain passed tests.²¹ In the method of agreement the investigator chooses cases with different characteristics and similar values on the dependent or independent variable, then asks if values on the other variable are similar across cases.

Controlled comparison is the most familiar case study method but the weakest. The method of difference is weak because in social sciences

¹⁷Theory-testing case studies are also known as "theory confirming" and "theory infirming" studies. "Comparative Politics and the Comparative Method," p. 692.

¹⁸See George and McKeown, "Case Studies and Theories," pp. 24-29, and works discussed in Collier, "Method," pp. 111-112 (his section on "Focus on Comparable Cases.")

¹⁹John Stuart Mill, *A System of Logic*, Books I-III, text ed. J.M. Robson (Toronto: University of Toronto Press, 1973) ("Of the Four Methods of Experimental Inquiry"), pp. 388-406. A discussion of Mill is George and McKeown, "Case Studies and Theories," pp. 26-28.

²⁰We also favor cases that have differing values on whichever variable is more easily measured to lower the labor exploring cases that turn out to have matching values on the study variable. Thus if we can judge value at a glance but must work to measure values on the DV, we select cases that differ in value on the IV (thus the requirements of the method of difference are met by our cases) and then dig out values on the DV. If value is easier to judge an values on the IV we select cases for difference in value on the DV.

²¹For a different view see Christopher H. Achen and Duncan Sissak, "Rational Deterrence Theory and Comparative Case Studies," *World Politics*, Vol. 41, No. 2 (January 1999), pp. 143-169 at 145-246. They write that case studies of deterrence "have failed when used for two tasks for which they are not suited--theory construction and theory verification." This failure lies "in the nature of the methods. The logic of comparative case studies inherently provides too little logical constraint to generate dependable theory and too little inferential constraint to permit trustworthy theory testing."

acteristics of paired cases are never nearly-identical (as the method of difference requires). The method of agreement is even weaker because paired cases usually deviate even further from having wholly different characteristics (as the method of agreement requires).²²

B. **Congruence procedure.**²³ The investigator explores the case looking for congruence or incongruence between values observed on the independent and dependent variable and values predicted by the test hypothesis. Two types of congruence procedure are used.

1. **Congruence procedure type 1: comparison to typical values.** The investigator observes within-case values on the IV and DV, and observes the world (without doing case studies) to ascertain values on the IV and DV that are typical in most other cases. The investigator then deduces from these observations and from the test theory expected relative within-case values for the IV and DV, and measures the congruence or incongruence between predicted and observed within-case values. For example, in a given case if the IV's value is above the typical norm, the value on the DV should also be above normal if the theory holds water. If the DV's value is below the typical norm, the value on the IV should also be below normal if the theory is valid. Etc.

How do we ascertain typical IV and DV values for other cases? Often the normal background levels of phenomena are known by common knowledge. Thus we know that Nazi Germany and Stalin's Soviet Union were far more murderous than typical modern industrial states without doing new studies to prove it, and we can safely compare their conduct to this typical conduct in a case study. Likewise, we do not need further study of general history to establish that elite belief that conquest was feasible was above the historical average in 1914 Europe. If common knowledge is thin or unreliable, however, research to establish typical values is necessary.

Congruence procedure type 1 works best if we select cases with extreme (very high or very low) values on the IV or DV. These cases are good test laboratories because theories make more unique and certain predictions about them. This allows stronger tests. In a case where a theory's IV is present in extreme abundance its effects (including both its intervening and dependent variables) should also be present in abundance. Likewise, if the IV is unusually scarce its effects should be more notable by their greater absence. If we observe these extreme results it is unlikely that they arise from measurement error, since only a large error would cause the observed result. The action of a third variable is also an unlikely cause of the observation, since it is unlikely that another cause operates strongly enough to produce the striking effect that the theory predicts. And any third variable

that was responsible would be present in abundance, hence it out against the case background, making it easy to spot. Hence easily rule out measurement error and omitted-variable explanation tests. (In other words, the tested prediction is quite unique test is strong.)

If we fail to observe the predicted result, it is less likely that an error or the countervailing effect of other variables caused the failure a large result was predicted, that result should have overpowered measurement errors or countervailing variables, visibly appearing. Moreover, a countervailing variable would probably need to be present and hence would be easy to spot. Hence a test failure in of a visible powerful countervailing variable casts large doubt on (In other words, the tested prediction is quite certain, hence strong.)

Hence both the passage and the flunking of tests provides decision in cases with extreme values on the IV. Passage strongly confirms hypothesis, a flunk strongly infirms it.

Type 1 congruence procedure is a close cousin of controlled (Both rest on cross-case comparisons, not within-case comparisons offer means to reduce the possibility that passed tests result from of third variables. They differ in the method of reducing this. Controlled comparison holds the case background constant, thus the variance of potentially perturbing third variables. Thus it is a range of variables that vary across cases, which lowers the number of perturbing variables. In contrast, if an "extreme value on study case selection method is used, Congruence type 1 reduces omitted problems by expanding the impact that omitted variables must produce the result predicted by the theory. This lowers the likelihood third variables have enough impact to produce this result, and a that these variables' necessarily extreme values will call attention selves if they do produce this result. Thus the number of potential variables is again reduced, this time by a different method.

²²Noting these and other difficulties with controlled comparison are George and McKewen, "Case Studies and Theories," p. 27; and Stanley Lieberman, "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases," *Social Forces*, Vol. 70, No. 2 (December 1991), pp. 307-320.

²³See George and McKewen, "Case Studies and Theories," pp. 29-34.

²⁴George describes congruence procedure (by which he means Type 1 congruence procedure) as a within-case because he argues that congruence or incongruence is established by deduction, not by comparison to other cases. Specifically, he argues that once we know the value on the IV we can deduce the expected value from the test theory. Then we assess the congruence or incongruence of this expectation with observed value the idea that expected within-case DV values are instead established by comparison to typical IV and DV values "Case Studies and Theory Development" (1982 Carnegie-Mellon paper), p. 14. However, it seems to me deductive exercise must rest on comparison to typical values in other cases, and on expectations about the cases are calibrated to these typical values. Hence it rests on cross-case comparison.

2. **Congruence procedure type 2: multiple within-case comparisons.** The investigator makes a number of paired observations of values on the IV and DV across a range of circumstances within a case. Then the investigator assesses whether these values co-vary in a manner that accords with the predictions of the test hypothesis.²⁵

Congruence procedure type 2 works best if we select cases with two characteristics: (1) many observations of values on the IV and DV are possible; and/or (2) values on the IV or DV vary sharply over time or across space (i.e., across region, institution, group, etc.) within the case.²⁶

Cases allowing many observations are better test laboratories because they allow more measures of congruence, and tests that rest on more measures are stronger.

Cases with large variation in values on the IV or DV are also good test laboratories because theories make more unique and certain predictions about these cases. For example, if values on a theory's IV vary sharply, values on its IntVs and DV should also vary sharply. This sharp variance on IntVs and DV is unlikely to arise from measurement error, since the error would need to be large and to gyrate in step with the IV--an unlikely combination. The action of a third variable is also an unlikely cause, since this would require a third variable that gyrates in step with 'A' and as markedly as 'A'--an unlikely possibility, and one that is easily assessed, since such a variable will leap out from the case. Hence we can more easily rule out measurement error and omitted-variable explanations for passed tests. For parallel reasons these explanations can also be ruled out for failed tests. As a result both the passage and the flunking of tests provides decisive evidence in cases with sharp variance on the IV. Passage strongly confirms the hypothesis, a flunk strongly infirms it.

Congruence procedure type 2 is a within-case case study, although it can shade into large-n analysis at some point, as the number of within-case observations grows larger and/or more uniform in character. For example, if the "case" is the 1994 US election, it is a bounded example of something more general, e.g., a parliamentary election in a democracy, hence it has aspects of a case. It also allows hundreds of observations that are scaled in a uniform way--e.g., the 435 US house races. These can be studied and statistically compared. Such a study has aspects of a congruence procedure and a large-n analysis.

Cases studies could, in principal, be hybrid combinations of congruence procedure types 1 and 2. An analyst could make many observations of a case compare these observations both to each other and to a typical average. The analyst also could select cases that offer many observations of IV and DV, or values on IV or DV, and large variance in these values.

Congruence procedures (both types) can be used to test a theory's explanatory hypotheses as well as its prime hypothesis.²⁷ To test explanatory hypothesis prefer cases that permit multiple measures of, and/or display extreme or varying values on, the causal and/or caused variables in the tested explanatory hypothesis.

C.

Process tracing.²⁸ The investigator explores the chain of events or the determining process by which initial case conditions are translated into outcomes. The cause-effect links that connect independent variable and outcomes are unwrapped and divided into smaller steps; then we look for observance of each step. Does the chain of events or decisionmaking process in the manner predicted by the theory? Specifically, do actors speak and as the theory predicts? Do they perceive and respond to stimuli in the predicted? Do the timing and details of their behavior match prediction the timing and details of other events that comprise the process that fit the initial conditions into outcomes match the theory's predictions? The tight fit between the theory's predictions about process and the actual determining process, the stronger the inference of validity.

Most theories make many predictions about causal process. Hence proceeding, like congruence procedure, allows the investigator to test many predictions within a single case. For example, a traceable process of causation hypothesis that "economic depression causes war" might be: deepening depression causes popular clamor for war, causing elites who represent those elites loudest to press for war, causing other elites to agree to war, causing This process gives rise to the following predictions of a case of a state suffers depression and fights a war: (1) as the depression deepens we should see growing arguments for war emerging in public debate; (2) war fever should develop faster and further among groups that groups most injured by the war; (3) elites that represent these groups should precede other elites in for war; (4) other elites should opt for war after these pressures are applied by diaries, private correspondence, and memoirs). We have one theory case but several predictions.

²⁵Alexander George, who coined the concept of congruence procedure, does not mention multiple within-case comparisons as a type of congruence procedure in his various writings on case studies, but his discussions of congruence procedure are consistent with the possibility of multiple observations and comparisons. See, e.g., George, "Case Studies and Theory Development" (1982 Carnegie-Mellon paper), pp. 13-15; George and McKeown, "Case Studies and Theories," pp. 29-34.

²⁶If we seek to explore the causes or effects of intervening variables (which constitute the "explanatory hypotheses" of the theory, see below), cases that allow many observations on IntV values and offer large variance on IntV values can be the most fruitful to study.

²⁷If the theory holds that "A → q → B", then "A → B" is its prime hypothesis. "A → q" and "q → B" are explanatory hypotheses. See Memo 1, "Hypotheses, Laws and Theories: A User's Guide," in this working paper.

²⁸See George and McKeown, "Case Studies and Theories," pp. 34-41; also King, Keohane and Verba, *Designing Inquiry*, pp. 226-228.