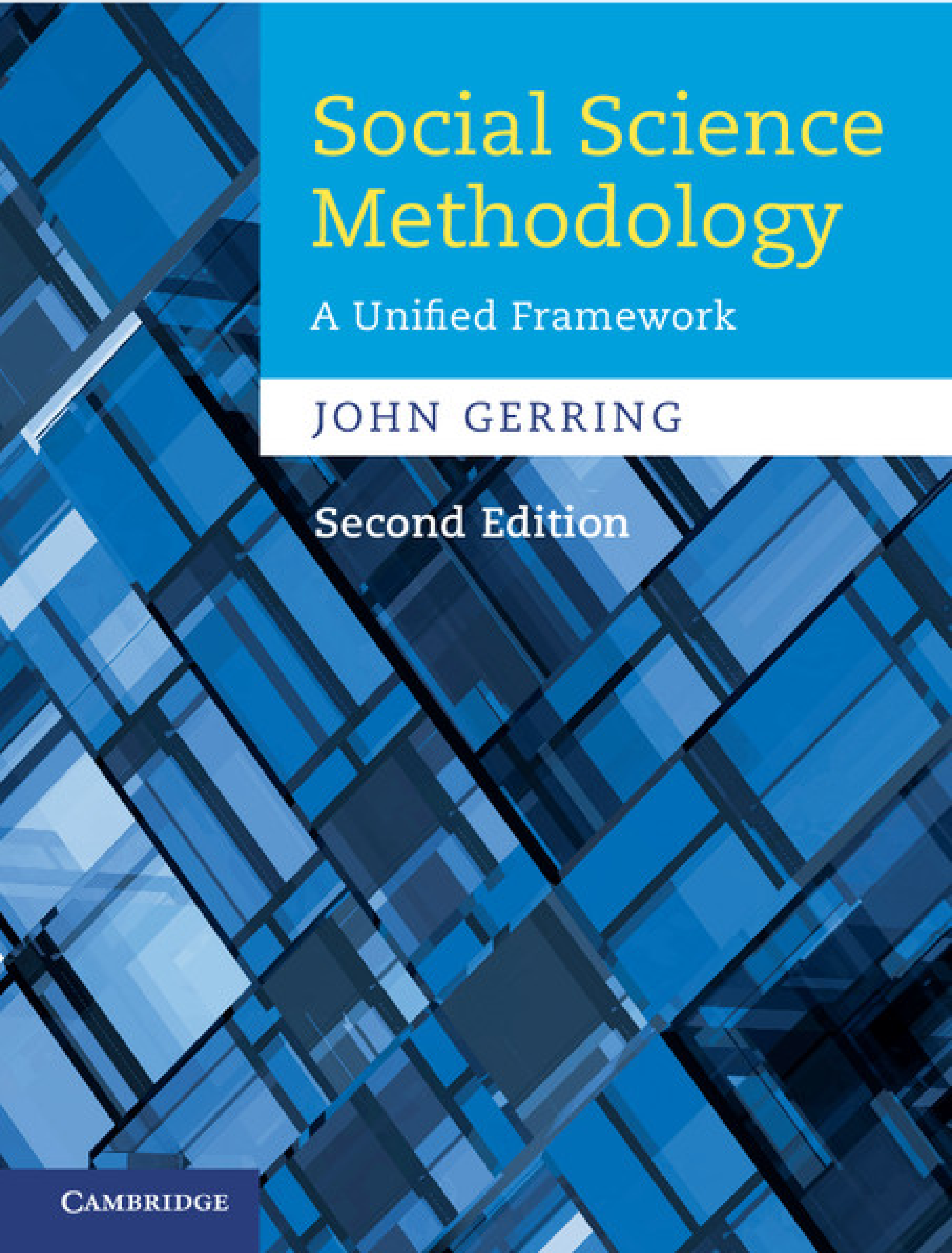
STRATEGIES FOR SOCIAL INQUIRY

Social Science Methodology

A Unified Framework

JOHN GERRING

Second Edition



CAMBRIDGE

Social Science Methodology

A Unified Framework

Second edition

John Gerring's exceptional textbook has been thoroughly revised in this second edition. It offers a one-volume introduction to social science methodology relevant to the disciplines of anthropology, economics, history, political science, psychology, and sociology. This new edition has been extensively developed with the introduction of new material and a thorough treatment of essential elements such as conceptualization, measurement, causality, and research design. It is written for students, long-time practitioners, and methodologists, and covers both qualitative and quantitative methods. It synthesizes the vast and diverse field of methodology in a way that is clear, concise, and comprehensive. While offering a handy overview of the subject, the book is also an argument about how we should conceptualize methodological problems. Thinking about methodology through this lens provides a new framework for understanding work in the social sciences.

John Gerring is Professor of Political Science at Boston University, where he teaches courses on methodology and comparative politics. He has published several books including *Case Study Research: Principles and Practices* (Cambridge University Press, 2007), and *A Centripetal Theory of Democratic Governance* (Cambridge University Press, 2008). He served as a fellow of the School of Social Science at the Institute for Advanced Study (Princeton, NJ), as a member of the National Academy of Sciences' Committee on the Evaluation of USAID Programs to Support the Development of Democracy, as President of the American Political Science Association's Organized Section on Qualitative and Multimethod Research, and was the recipient of a grant from the National Science Foundation to collect historical data related to colonialism and long-term development. He is currently a fellow at the Kellogg Institute for International Affairs, University of Notre Dame (2011–12).

Strategies for Social Inquiry

Social Science Methodology: A Unified Framework (second edition)

This new book series presents texts on a wide range of issues bearing upon the practice of social inquiry. Strategies are construed broadly to embrace the full spectrum of approaches to analysis, as well as relevant issues in philosophy of social science.

Forthcoming Titles

Michael Coppedge, *Approaching Democracy: Theory and Methods in Comparative Politics*

Thad Dunning, *Natural Experiments in the Social Sciences*

Diana Kapiszewski, Lauren M. MacLean and Benjamin L. Read, *Field Research in Political Science*

Jason Seawright, *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*

Carsten Q. Schneider and Claudius Wagemann, *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*

Social Science Methodology

A Unified Framework

Second edition

John Gerring

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521132770

First published 2012

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-11504-9 Hardback
ISBN 978-0-521-13277-0 Paperback

Additional resources for this publication at www.cambridge.org/gerring

There is no royal road to science, and only those who do not dread the fatiguing climb of its steep paths have a chance of gaining its luminous summits.

Karl Marx, "Preface to the French Edition," *Capital* (299),
quoted in Levi (1999: 171)

To have mastered "method" and "theory" is to have become a self-conscious thinker, a man at work and aware of the assumptions and the implications of whatever he is about. To be mastered by "method" or "theory" is simply to be kept from working, from trying, that is, to find out about something that is going on in the world. Without insight into the way the craft is carried on, the results of study are infirm; without a determination that study shall come to significant results, all method is meaningless pretense.

C. Wright Mills, *The Sociological Imagination* (1959: 120–121)

Surely, in a world which stands upon the threshold of the chemistry of the atom, which is only beginning to fathom the mystery of interstellar space, in this poor world of ours which, however justifiably proud of its science, has created so little happiness for itself, the tedious minutiae of historical erudition, easily capable of consuming a whole lifetime, would deserve condemnation as an absurd waste of energy, bordering on the criminal, were they to end merely by coating one of our diversions with a thin veneer of truth. Either all minds capable of better employment must be dissuaded from the practice of history, or history must prove its legitimacy as a form of knowledge. But here a new question arises. What is it, exactly, that constitutes the legitimacy of an intellectual endeavor?

Marc Bloch, *The Historian's Craft* ([1941] 1953: 9)

Contents

Detailed table of contents

Part II Description

4 Analyses

But is it true?

Aaron Wildavsky[1]

Having discussed the formal (super-empirical) criteria of a good argument, we turn now to the empirical portion of social science research, the hoped-for encounter with reality.[2] This stage may be referred to variously as analysis, assessment, corroboration, demonstration, empirics, evaluation, methods, proof, or testing. (While acknowledging the subtle differences among these terms, I shall treat them as part of the same overall enterprise.)

Of course, the distinction between theory formation and theory-testing is never clear and bright. As is the case everywhere in social science, tasks intermingle. One cannot form an argument without considering the empirical problem of how to appraise it, and vice versa. Moreover, the task of (dis)-confirming theories is intimately conjoined with the task of forming theories. As Paul Samuelson notes, "It takes a theory to kill a theory."[3]

Yet in coming to grips with the complex process of social science it is essential to distinguish between the formal properties of an argument and the methods by which that argument might be assessed. *What are you arguing?* and *Is it true?* are logically distinct questions, calling forth different criteria of adequacy.[4] Moreover, there are good methodological reasons to respect the separation between theory and analysis (see "Partition" below). We now proceed from the former to the latter.

Of course, not all hypotheses require explicit attention to methods of appraisal. Many hypotheses need not be formally tested at all, for they are already self-evident (e.g., "civil war is dislocating"), or are insufficiently important to justify the investment of time and energy that a formal analysis would require (e.g., "lifeguard training programs have positive effects on the probability of

[1] Wildavsky (1995).

[2] Scientific realists recognize an analogous distinction between the super-empirical and empirical elements of a theory (Hitchcock 2003: 217).

[3] Quoted in Rosenbaum (2010: 95).

[4] Bhaskar ([1975] 1978: 171); Bunge (1963: 45); Hoover (2001: 22).

marriage and child-bearing among program participants"). Our motivation here is centered on arguments that are important enough to submit to a formal testing procedure and complex enough, in terms of potential threats to validity, to worry about the niceties of research design. Methodology kicks in where common sense falls short.

Definitions

A standard empirical analysis involves a number of components, which must be clarified before we continue. Much of this vocabulary is borrowed from survey research; nonetheless, the concepts are helpful in all styles of research, quantitative or qualitative.

A *population* is the universe of phenomena that a hypothesis seeks to describe or explain. It remains unstudied, or is studied only in a very informal manner, for example, through the secondary literature. Sometimes, it is important to distinguish between a population from which a sample is drawn (and which it presumably represents) and a larger, more hypothetical population that the sample may or may not represent, but which nonetheless defines the scope-conditions of the argument.

The *sample* refers to the evidence that will be subjected to direct examination. It is composed of *units* or *cases*: bounded entities such as individuals (subjects), organizations, communities, or nation-states, which may be observed spatially and/or temporally (through time). (The terms *unit* and *case* are more or less equivalent. The only difference is that while a unit is bounded spatially, a case may also have implicit or explicit temporal boundaries.[5])

Typically, the sample is smaller than the population; hence, the notion of *sampling* from a population. (Note, however, that my use of the term sample does not necessarily mean that cases under study – the sample – have been randomly chosen from a known population.) Occasionally, one is able to include the entire population in a sample – a *census*.

The *observations* taken from units at particular points (or periods) in time compose the pieces of evidence presumed to be relevant to a descriptive or causal proposition. Collectively, the observations in a study comprise a study's sample. Each observation should record values for all relevant *variables* across each unit at a particular point (or period) in time. In causal analysis, this

[5] For further discussion see Gerring (2007).

includes *X* (the causal factor of theoretical interest) and *Y* (the outcome of interest), along with any other variables deemed essential for the analysis.

In matrix format, an observation is usually represented as a row and the total number of observations (rows) in a sample as "*N*." Confusingly, *N* also sometimes refers to the number of units or cases, which may be quite different from the number of observations. Varying usages are usually clear from the context.

A final concept, the data *cell*, is useful when one wishes to refer to the data pertaining to a particular unit at one point in time along only one dimension. Although the term is not commonly employed, it is sometimes essential. Consider that an observation consists of at least two cells in any causal analysis: the cell representing the value for *X* and the cell representing the value for *Y*. Sometimes, one needs to distinguish between them.

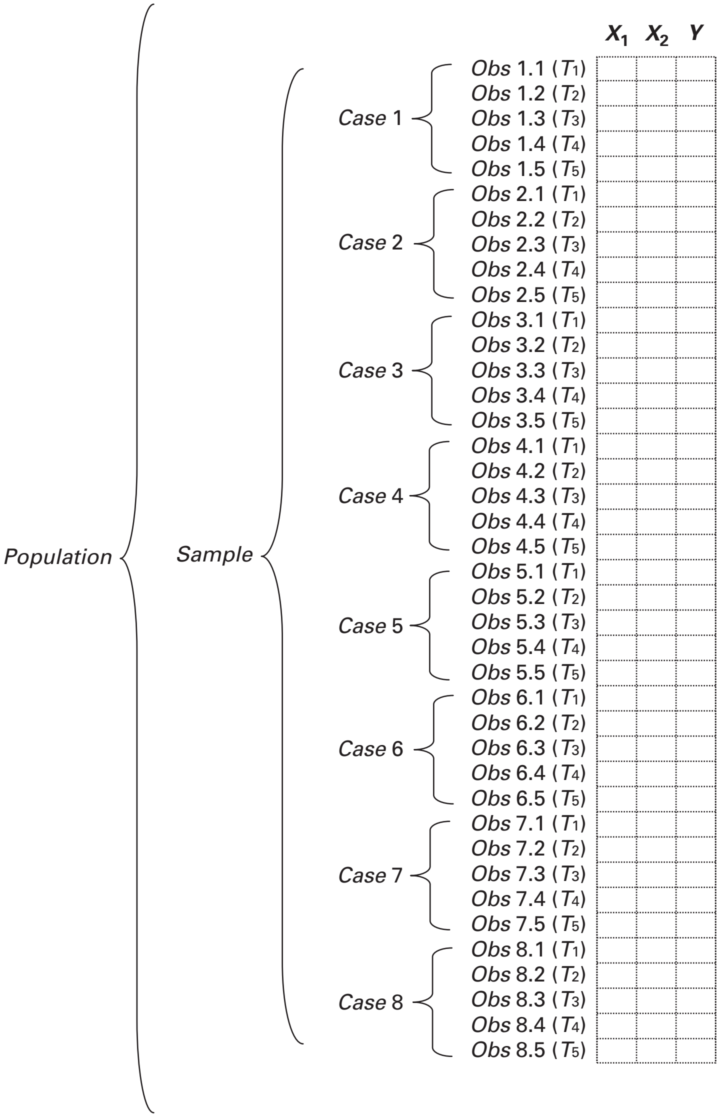
These interrelated concepts are illustrated in Figure 4.1, where we can see a fairly typical time-series cross-section research design in a rectangular dataset (matrix) format. Here, observations are represented as rows, variables as columns, and cells as their intersection. Note that cells are nested within observations, observations are nested within units (aka cases), units are nested within the sample, and the sample is nested within the population.

Hypothetically, let us imagine that the population of the inference includes all US schools and the sample consists of eight schools, observed annually for five years, yielding a sample of forty observations (*N*=40). The *units of analysis* (the type of phenomena treated as observations in an analysis) in this hypothetical example are school-years.

If the research design had been purely cross-sectional, only one observation would be taken from each unit, and the units of analysis would consist of schools rather than school-years, and the total number of observations would be eight (*N*=8). In this context, the number of units is equal to the number of observations and the distinction between unit and observation is lost.

If the research design is purely temporal the sample would be composed of one unit, observed through time. If the sample period is five years and observations are taken annually, the total number of observations is five (*N*=5). Here, the units of analysis are again school-years, as in the first example.

All these terms are slippery insofar as they depend for their meaning on a particular proposition and a corresponding research design. Any changes in that proposition may affect the sort of phenomena that are classified as observations and units, not to mention the composition of the sample and the population. Thus, an investigation of school vouchers might begin by identifying *schools* as the principal unit of analysis, but then shift to a lower level of analysis



Population = indeterminate; Cases/units = 8; Sample/observations = 40; Cells = 120; Time (*T*) = 1–5; Variables = 3.

Figure 4.1 Time-series cross-section dataset

(e.g., *students*), or a higher level of analysis (e.g., *school districts*) at different points in the study. Sometimes, different levels of analysis (e.g., students, schools, and school districts) are combined. This is common in *case study* work and is the defining feature of *hierarchical* (*multi-level*) statistical models.

Complicating matters further, the precise boundaries of a research design often remain ambiguous. This is because a subject is usually interrogated in a variety of ways during the course of a study. For example, key variables may change (perhaps to capture a different dimension or an alternative operationalization of a complex concept), the units of analysis may change (moving up or down in levels of analysis), the focus may change (from the main hypothesis to adjunct hypotheses or causal mechanisms), the sample may change, and different kinds of observations may be enlisted. These are just a few of the variations in method that typically co-habit in a single study. Each of these alterations may be considered as distinct research designs or as variations on a single research design. Likewise, they may be described as replications, robustness tests, or multimethod research (as discussed in later chapters). Thus, it becomes rather difficult to say what a given study's research design is, or how many there are, without making some rather arbitrary decisions about what lies in, and out of, the scope of this ambient concept. I shall leave this matter open because I do not think it can be easily settled. Perhaps it is not essential. The proviso is that writers must be clear about what they mean by "research design" in a given context.

Research design versus data analysis

Traditionally, one distinguishes between two stages of the testing process. *Research design* refers to the selection and arrangement of evidence.[6] *Data analysis* refers to the analysis of data once it is collected.

In an experiment, these stages are clearly separable: research design precedes data analysis. One is *ex ante*, the other *ex post*. (Of course, in successive cycles of research this line becomes blurred.) In observational research, the two stages are usually intermixed. Because much of this book is focused on observational techniques, the reader should be prepared for some slippage across these two concepts. Still, the distinction is consequential.

An older tradition of social science methodology focuses on reaching inferences about a phenomenon based on whatever data is at hand. The methodologist's job begins once the evidence is in. This is the "data analysis" approach to methodology that underlies most econometrics texts. Textbooks in this genre include discussions of statistical inference and of various classes of estimators

[6] An experimentally based understanding of design refers to "all contemplating, collecting, organizing, and analyzing of data that takes place prior to seeing any outcome data" (Rubin 2008: 810). This seems to narrow for present purposes, since in observational research the selection of a research site often depends on an initial consideration of "outcome" data. My understanding of design encompasses all factors that might (legitimately) impact the choice of observations to be studied.

employed for descriptive and causal inference (e.g., correlation, difference of means, regression, matching, randomization inference, Bayesian versus frequentist approaches), along with the assumptions each method invokes.[7]

Useful though such techniques are, it is important to remember that the contribution of advanced statistical protocols is focused largely on shortcomings of design. Econometrics is the *deus ex machina* hauled onto the stage to rectify problems of measurement error, ambiguous causal factors, insufficient variation along key parameters, insufficient observations, incomparabilities across comparison cases, biased samples, and other issues that we will shortly discuss. From this perspective, it seems appropriate to conclude that matters of design are primary, and matters of data analysis secondary – both sequentially and methodologically. "Design trumps analysis," in the words of Donald Rubin.[8] And from this perspective it follows that the methodologist's job begins at the front-end – the research design phase of a project.

Indeed, there is often not much one can do to rectify problems of design once the data is in. For those who are fond of medical analogies, the research design approach to methodology might be compared with the preventive approach to medicine, that is, how to avoid contracting illness, while the data analysis approach to methodology is akin to emergency care, that is, how to restore a patient who is already failing.

Sometimes, ingenious *ex post* statistical adjustments are successful. Yet there is increasing skepticism about our capacity to correct research design flaws at the post-research phase. The old adage, "garbage in, garbage out," is still true, despite many advances in the field of statistics. Richard Berk comments:

> One cannot repair a weak research design with a strong data analysis. Almost inevitably what seems too good to be true is, and one is simply substituting untestable assumptions for the information one does not have.[9]

Indeed, the most worrying point of all is we usually cannot tell whether statistical corrections have achieved their intended purpose, for example, whether a two-stage approach to modeling selection bias has actually provided a correct and unbiased estimate of *X*'s effect on *Y*. As Berk points out, this is because the assumptions required to conduct statistical protocols are often not directly testable; they hinge on *a priori* ("ontological") assumptions about the nature of the data-generating process. Reviewing the field of regression-based causal

[7] For example, Greene (2002).

[8] Rubin (2008). See also Angrist and Pischke (2010); Bowers and Panagopoulos (2009); King, Keohane, and Verba (1995); Rosenbaum (1999, 2010); Sekhon (2009); Shadish and Cook (1999: 294).

[9] Berk (1991: 316).

inference, David Freedman states baldly, “I see *no* cases in which regression equations, let alone the more complex methods, have succeeded as engines for discovering causal relationships.”[10] While this conclusion seems a tad extreme, one is rightly cautioned to regard statistically based causal inferences with skepticism. Always, they rest on assumptions about the data-generation process, that is, on matters of research design.

Thus, although I do not wish to downplay the importance of data analysis, I do wish to stake a claim for the primacy of design – especially in causal analysis but also in descriptive analysis. The design components of research are general in purview; any attempt to disentangle empirical relationships must wrestle with them. Moreover, this perspective on methodology is often insightful. It clarifies the obstacles facing the social sciences and elucidates a range of possible solutions.

Finally, the design aspects of social science research are under-appreciated. Indeed, the only regions of social science where issues of design are granted primacy are those where experimental methods are employed. In light of this, it seems arguable that the way forward for social science is to be found in well-crafted research designs rather than in the development of new estimators. Borrowing from Paul Rosenbaum, our motto will be “choice as an alternative to [statistical] control.”[11]

Accordingly, the following chapters include little discussion of statistics *except* as the latter bear upon matters of research design. This means that statistical methods closely associated with specific research designs, such as regression discontinuity and instrumental variables, will be discussed (Chapter 10), but not statistical methods that are general in employment, such as regression or matching.

Criteria

With these terms and perspectives clarified, we can now proceed to the main business at hand. What is it that qualifies a research design (and

[10] Freedman (1997: 114; emphasis added). On the problems of statistical inference based on observational data, and the corresponding importance of research design, see Berk (2004); Brady and Collier (2004); Clogg and Haritou (1997); Freedman (1991, 2008, 2010); Gerber, Green, and Kaplan (2004); Gigerenzer (2004); Heckman (2008: 3); Kittel (2006); Longford (2005); Pearl (2009b: 40, 332); Robins and Wasserman (1999); Rodrik (2005); Rosenbaum (1999, 2005); Seawright (2010); Summers (1991). Various studies comparing analyses of the same phenomenon with experimental and nonexperimental data show significant disparities in results, offering direct evidence that observational research is flawed (e.g., Benson and Hartz 2000; Friedlander and Robins 1995; Glazerman, Levy, and Myers 2003; LaLonde 1986). Cook, Shaddish, and Wong (2008) offer a more optimistic appraisal.

[11] Rosenbaum (1999).

Table 4.1 Analysis: general criteria

1. **Accuracy**
   Are the results (a) valid, (b) precise (reliable), and (c) accompanied by an estimate of uncertainty (confidence, probability) with respect to (d) the chosen sample (internal validity) and (e) the population of interest (external validity, aka generalizability)?
2. **Sampling**
   Are the chosen observations (a) representative of the intended population, (b) sufficiently large in number, and (c) at the principal level of analysis?
3. **Cumulation**
   (a) Is the research design standardized with other similar research on the topic? (b) Does it replicate extant findings and facilitate future replications by other scholars? (c) Are procedures transparent?
4. **Theoretical fit**
   (a) Does the research design provide an appropriate test for the inference (construct validity)? (b) Is the test easy or hard (severity)? (c) Is the test segregated from the argument under investigation (partition)?

corresponding data analysis) as satisfactory? What is a good empirical analysis?

I will argue that criteria applicable to social science analyses may be fruitfully divided into four fundamental areas: *accuracy* (validity, precision, and uncertainty); *sampling* (representativeness, sample size, level of analysis); *cumulation* (standardization, replication, transparency); and *theoretical fit* (partition, construct validity, difficulty).

These criteria, summarized in Table 4.1, are regarded as generic, which is to say they apply to all approaches. No method – whether descriptive or causal, qualitative or quantitative, experimental or observational – is exempt. To be sure, each study is apt to prioritize certain criteria over others. And occasionally, criteria may be legitimately ignored if they have been effectively established by other studies. In this respect, it is difficult to evaluate a given work in isolation from the field of studies in which it is situated. But the larger and more important claim remains: the criteria listed in Table 4.1 are broadly applicable wherever empirical questions of social science are in play.

Accuracy

The overall objective of empirical research is to accurately test an argument. Accuracy may be understood as having two dimensions: *validity* and *precision*, each with an associated level of *uncertainty*.

These notions are typically applied to the estimate that results from an empirical analysis (i.e., to the finding). However, they may also be applied to the research design and technique of data analysis by which that estimate is obtained. Indeed, the various phases of research are *all* subject to demands for validity and precision, and each is associated with a level of uncertainty. Thus, when speaking of these goals we shall speak of them applying across various tasks associated with the general task of theory appraisal.

Other criteria, discussed in succeeding sections of this chapter and in subsequent chapters, usually aim in one way or another to bolster the accuracy of an analysis, and in this respect may be viewed as ancillary to the fundamental goals of validity and precision.

Finally, a distinction will be introduced between the chosen sample and a larger population of theoretical interest. The former is understood as an issue of *internal validity* and the latter as an issue of *external validity*.

Validity, precision, uncertainty

Scholars often distinguish between the *validity* of a test and its *precision* (reliability). If an inference were to be tested repeatedly, the closeness of these results (on average) to the true value would capture the validity of the test. The closeness of these test results *to each other* would capture the precision of the test.

This contrast is best illuminated by illustration. Let us represent the object of interest (in its true, ontological reality) by a dark circle, and various attempts to measure that object by points. With this schema, three tests are compared in Figure 4.2. The first is reliable but not valid, as the points cluster closely together but are distant from the true center. The second is valid but not reliable, as the points are dispersed but are clustered around the true center. The third is both reliable and valid.

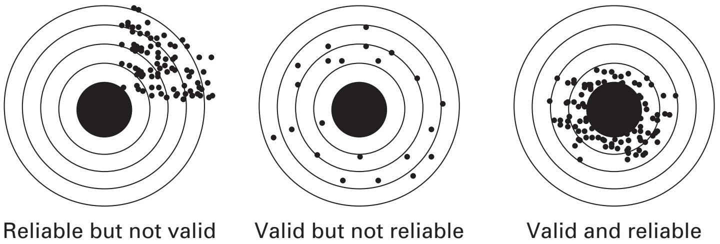


Figure 4.2 Reliability (precision) and validity

These concepts apply equally to the descriptive task of measurement (Chapter 7), as well as to the task of estimating causal effects (Chapters 9, 10, and 11). There is a slight alteration of vocabulary, insofar as the precision of a measurement is usually referred to as a question of *reliability* (rather than precision). But the basic ideas are the same across contexts.

Note that precision is also a criterion of an argument (Chapter 3). Here, however, we are concerned with the precision of a test, not the precision of the proposition that is being tested.

Now, let us explore these issues in greater detail.

A problem of validity may be expressed as a problem of *systematic error* or *bias*. Of course, it depends upon assumptions about the true reality, which may not be directly apprehensible. In some circumstances, it is possible to gauge the validity of a statistical model through *Monte Carlo* simulations.[12] But usually issues of validity are assessed in a more speculative manner. If there is recognizable bias, or potential bias, in some aspect of the research design we say that there is a problem of validity – even though we cannot know for sure.

Precision, we have said, refers to the consistency of a finding across repeated tests, and is thus a large-sample property. If iterated tests demonstrate the same result (more or less), the procedure is deemed to be precise. The *variance* across these results provides an empirical measure of the degree of precision thereby attained. If there is no opportunity to compare multiple iterations of a single research design (if the research is qualitative in nature), then the variance remains a theoretical property – though no less important for being so. Many factors may affect the relative precision of a test, including measurement error, the variability of the phenomena under study, and the size of a sample. Since precision is about variance, not validity, all such errors are regarded as stochastic (random), aka noise.

Implicit in the notion of validity is the concept of *uncertainty*. Any assertion about the world is associated with a level of confidence, or probability; for all empirical knowledge is to some extent uncertain. There is always a problem of inference, even if the degree of uncertainty is judged to be quite small. This uncertainty may stem from problems of concept formation (Chapter 5), measurement (Chapter 7), sampling (discussed below), and/or various issues associated with causal inference (Part III). It depends, obviously, on the argument in question.

It has been alleged that “perhaps the single most serious problem with qualitative research . . . is the pervasive failure to provide reasonable estimates

[12] Mooney (1997).

of the uncertainty of the investigator's inferences."[13] I have no doubt that there is some truth to this assertion, though qualitative scholars have worked hard to resolve it.

By contrast, quantitative methods generate estimates of uncertainty as a routine element of the analysis. Certain aspects of uncertainty can be captured in a statistic such as a confidence interval and associated *p* value, which measures the probability of a hypothesis relative to some null hypothesis. Here, the concepts of precision and uncertainty are merged in a single statistic. To be sure, these statistics are based on sampling variability and thus take no account of other threats to inference. Bayesian approaches are broader in reach, incorporating subjective knowledge about a subject. It is in this spirit that I propose an encompassing approach to the estimation of uncertainty, one that combines information drawn from large-sample methods of inference (wherever samples are large enough to permit this) with qualitative knowledge about additional threats to inference. Estimating the uncertainty of a particular finding is not easy. But it is essential.

Internal/external validity

Conventionally, one analyzes questions of validity, precision, and uncertainty at two levels. First, there is the question of whether a finding is true for the chosen sample – an issue of *internal* validity. Second, there is the question of how this finding might be generalized to a broader population of cases – an issue of *external* validity. Note that although this is phrased in terms of validity the same questions arise with respect to precision; I shall therefore assume that both are inferred when one utters the phrase "internal validity" or "external validity."

A study may be valid internally but not externally (beyond the chosen sample or research site). Likewise, the internal validity of a study may be questionable, while its claim to external validity – if true for the sample – is strong. Of course, the issue of external validity rests in some important sense on a study's internal validity. The greater our confidence about a finding in context *A* (the chosen research site), the greater our confidence about that finding in context *B* (somewhere in the larger population of interest). By the same token, if one is not confident about a result within a studied domain one is even less confident about extending that result to a larger domain.

[13] King, Keohane, and Verba (1994: 32).

The internal/validity distinction is crucial to virtually every methodological discussion, even though the dichotomy is not always crystal clear. As an example, consider a hypothetical study of a school district in the state of New York that rests on a sample of students drawn from that district, but purports to elucidate features of all schools within the state. This presents three potential levels of validity: (1) the sample of students; (2) the school district; and (3) schools throughout the state (across multiple districts). Internal validity may refer to (1) or (2), while external validity may refer to (2) or (3).

In this light, the issue of internal/external validity is perhaps more correctly articulated as *degrees of generalizability*. Just as arguments aim to generalize, so do research designs. Some do so more successfully, and more extensively (across a broader population) than others. In this vein, it is sometimes helpful to recognize concentric circles surrounding the sample that has been studied. Typically, the confidence with which one extrapolates results obtained from a given sample decreases as the size of the circle expands. Returning to the example above, let us consider six possible tiers of validity: (1) the sample of students; (2) the school district; (3) schools throughout the state (across multiple districts); (4) schools in other states; (5) schools in other countries in the OECD; and (6) schools elsewhere in the world. Each succeeding claim to validity seems less likely, but none is wholly implausible. And from this perspective there is no clear demarcation between internal and external. Or perhaps there is a fairly clear demarcation between internal and external, but there are multiple spheres of external validity.

For heuristic purposes, subsequent discussion will assume that there is one context for a study that is appropriately labeled "internal" and another that is appropriately labeled "external." But readers should bear in mind the attendant complexities.

In rare instances, the distinction between internal and external validity disappears because the entire population of an inference is directly studied. Here, the sample *is* the population. Even so, there is room for skepticism about exhaustive sampling procedures (a *census*). Since most social science theories are not limited to the past, the future provides a potential source for out-of-sample testing. This means that even if all available examples that fall into the domain of a subject are studied one may still be theoretically motivated to understand in a much larger – as yet unfathomable – population.

Conceptually, one may also recognize a distinction between cases that actually exist and those that could have existed (in the past). Thus, if I am studying the relationship between economic development and democracy among nation-states in the modern era I might consider even a comprehensive sample – including all nation-states since 1800 – to be a sample of all the nation-states that could have existed during that time period. From this perspective, there is always a larger population that cannot be directly studied.

Note that the distinction between internal and external validity is grounded in a distinction between what has been directly studied and what has not been directly studied. This means that the issue of external validity cannot be tested, by definition. It rests at the level of assumption. (Of course, it may be tested by some future study.) The question arises, on what (speculative) basis does one judge a study's external validity?

The most obvious criterion is the *representativeness* of the sample, as discussed below. A more subtle issue – relevant only to causal analysis – is the *scalability* of the treatment, as discussed in Chapter 9.

Sampling

The selection of units and observations for analysis is critical to any descriptive or causal analysis. Three objectives pertain broadly to this task: *representativeness*, *size*, and *level of analysis*. In constructing a sample one should aim to be representative of a broader population, to include sufficient observations to assure precision and leverage in the analysis, and to use cases that lie at the same level of analysis as the primary inference.

Representativeness

The most important ground for drawing conclusions about the external validity of a proposition is the *representativeness* of a chosen sample. Is the sample similar to the population with respect to the hypothesis that is being tested? If, for example, the hypothesis is causal, then the question is whether the relationship of X to Y is similar in the sample and in the population. Are we entitled to generalize from a given sample to a larger universe of cases?

In the case of voucher research, one must wonder whether the students, schools, and school districts (along with whatever additional features of the research site may be relevant to the inference) chosen for analysis are representative of a larger population of students, schools, and districts. And, if so, what is that larger population? Does it consist of all students and schools across the United States, or across the world? Does it consist of a smaller population of students who are willing to volunteer for such programs? These are critical questions. Unfortunately, they are often difficult to answer in a definitive fashion for the reasons already discussed.

The best way to obtain a representative sample is to sample randomly from a larger population. There are many techniques for doing so (much depends upon the character of that larger population, the methods at one's disposal for sampling from it, and the inference one wishes to estimate). But the basic idea is that each unit or observation within the population should have an equal chance of being chosen for the sample. An advantage of this approach is that one can estimate sampling variability (from sample to sample), thus providing estimates of precision to accompany whatever inferences one wishes to draw.[14]

Unfortunately, it is not possible to apply methods of random sampling to many research problems. Voucher studies, for example, depend upon the willingness of school districts to implement their protocols – a rare occurrence. As such, the sample of school districts studied by researchers is not likely to be drawn randomly from the general population.

Even where random sampling procedures are feasible, they are not always methodologically defensible. If the sample under study is very small – say, a single case or a handful of cases – it does not make sense to draw randomly from a large population. While the chosen sample will be representative of the population *on average*, any given sample (of one or several) is quite likely to lie far from the mean (along whatever dimensions are relevant to the question under study). Consequently, case-study research generally relies on purposive (non-probability) case-selection strategies, reviewed elsewhere.[15]

Wherever random sampling techniques are inapplicable, researchers must struggle to define the representativeness of a sample, and hence the plausible generalizability of results based on that sample. This is true regardless of whether the sample is very small (i.e., a case-study format) or very large.

Before concluding this section it is important to remind ourselves that the goal driving the selection of a set of cases is not simply to assure representativeness (and, hence, external validity). It is also, and perhaps more importantly, to achieve internal validity. Frequently, these two objectives conflict. For example, researchers often find themselves in situations where they can craft an experiment with a nonrandom sample *or* conduct a nonexperimental study with a random sample. Usually, they opt for the former approach, signifying that they place greater priority on internal validity than on external validity. But in some situations one can imagine making the opposite choice.

[14] Weisberg (2005).

[15] See Gerring (2007: ch. 5).

Size (*N*)

More observations are better than fewer; hence, a larger "*N*" (sample size) is superior to a smaller *N*, all other things being equal. (*N* may be understood as standardized "dataset" observations or as irregular "causal-process" observations, a distinction introduced in Chapter 11.) This is fairly commonsensical. All one is saying, in effect, is that the more evidence one can muster for a given proposition, the stronger the inference will be. Indeed, the same logic that compels us to provide empirical support for our beliefs also motivates us to accumulate multiple observations. The plural of "anecdote" is "data," as the old saw goes.

Suppose one is trying to figure out the effect of vouchers on school performance, but one has available information for only one student or one school. Under the circumstances, it will probably be difficult to reach any firm conclusions about the causal inference at issue. Of course, one observation is a lot better than none. Indeed, it is a quantum leap, since the absence of observations means that there is no empirical support whatsoever for a proposition. Yet empirical research with only one observation is also highly indeterminate, and apt to be consistent with a wide variety of competing hypotheses. Consider a scatter-plot graph of *X* and *Y* with only one data point. Through this point, Harry Eckstein observes, "an infinite number of curves or lines can be drawn."[16] In other words, one cannot know from this information alone what the true slope of the relationship between *X* and *Y* might be, and whether the relationship is in fact causal (a slope different from 0). The more observations one has, the less indeterminacy there is, and the more precision, with respect to *X*'s probable relationship to *Y*. Note that with a small sample, results are necessarily contingent upon the (perhaps peculiar) characteristics of the several chosen observations. Conclusions about a broader population are hazardous when one considers the many opportunities for error and the highly stochastic nature of most social phenomena.

A large sample of observations also helps with other tasks involved in causal assessment. It may assist in formulating a hypothesis – clarifying a positive and negative outcome, a set of cases which the proposition is intended to explain (the population), and operational definitions of the foregoing. All these issues become apparent in the process of coding observations, wherever there are multiple observations. But if there is only one observation, or multiple observations drawn from a single unit, these tasks often remain ambiguous. The

[16] Eckstein (1975: 113).

problem is that with a narrow empirical ambit the researcher is faced with an over-abundance of ways to operationalize a given hypothesis. School performance – the main outcome at issue in our vouchers example – could be measured by any observable feature in a given school. By contrast, where multiple schools are being observed the range of possible outcome measures is inevitably narrowed (by virtue of the paucity of data or costliness of tracking myriad indicators). Likewise, it will be necessary to stipulate in more certain terms how "success" will be defined – for the comparisons across schools must be explicit. The process of measurement across multiple observations forces one to come to terms with issues that might otherwise remain latent, and ambiguous.

One exception to the large-*N* criterion concerns an empirical study whose purpose is to disprove a causal or descriptive law (an invariant, "deterministic" proposition). As long as the observed pattern contradicts the hypothesis, a law may be disproven with a single observation.[17]

In all other settings, a larger sample is advisable – with the usual *ceteris paribus* caveat. Thus, if increasing the size of a sample decreases the representativeness of the sample one might decide that it is not worth the sacrifice: a smaller, more representative sample is superior. If one is limited by time or logistical constraints to study *either* a large sample of cross-case observations *or* a smaller sample of within-case observations one might decide that the latter offer stronger grounds for causal inference (for any of the reasons to be discussed in Part III). In short, there are many situations in which a smaller sample is preferred over a larger one. However, the reasons for this preference lie in other criteria. That is why it is still correct to view the size of a sample as a fundamental (*ceteris paribus*) criterion of social science.

Before concluding this section I must briefly mention the problem of *missing data*, as it intersects both sample representativeness and sample size. Usually, what is meant by missing data is that a sample lacks observations for some units that should (by some principle of selection, random or otherwise) be included. If the pattern of missing-ness is systematic, then the sample will be biased. If, on the other hand, it can be determined that the pattern of missing data is random, then the sample will be smaller than it should, but still perhaps representative (or at least as representative as it would have been without the missing data). A potential solution, if patterns of missing-ness are fairly predictable (using known data points) and the number of missing data points (relative to the total sample) is not too large, is to impute missing data.[18] In other situations, it may be feasible to generate a simple decision rule for establishing a "best guess" for

[17] Dion (1998). [18] Allison (2002).

missing data points, without a formal statistical model. In any case, patterns of missing-ness must be reckoned with. A sample of 1,000 with missing data is not the same as a sample of 1,000 with no missing data. When one considers the problem of sample size one must wrestle with the completeness of the observations comprising the sample.

Level of analysis

Observations are most helpful in elucidating relationships when situated at the same level of analysis as the main hypothesis.[19] If the central hypothesis concerns the behavior of schools, then schools should, ideally, comprise the principal unit of analysis in the research design. If the hypothesis is centered on the behavior of individuals, then individuals should be the principal unit of analysis. And so forth.

One often faces difficulties if one attempts to explain the activity of a particular kind of unit by examining units at a higher, or lower, level of analysis. Suppose, for example, that one is interested in explaining the behavior of schools but has data only at the district level (an aggregation of schools). This is a common situation, but not an enviable one, for one must infer the behavior of schools from the behavior of school districts (raising a problem of estimation known as *ecological inference*).[20]

If, conversely, one has data at a lower level of analysis (for example, for students) then one faces a similar problem in the reverse direction: one must infer upward, as it were, from students to schools. This species of inference is also problematic. Sometimes, macro-level phenomena do not reflect observable phenomena at the micro-level, introducing a problem of *reductionism* (aka the *fallacy of nonequivalence*). Granted, knowing something about the response of students to a stimulus may be extremely helpful in understanding the response of schools. Indeed, it may be crucial to demonstrating the causal mechanism(s) at work. This is why case-study research, which typically invokes data lying at a lower level of analysis, is often employed. However, in proving the existence of a causal effect it is important also to muster evidence at the principal unit of analysis (as defined by the proposition). In this context, student-level data will be most useful if it can be aggregated across schools. And for purposes of estimating the *size* of a causal effect, along with some level of *precision/uncertainty*, observations drawn from the principal level of analysis are essential.

[19] Lieberson (1985: ch. 5). [20] Achen and Shively (1995).

While the level-of-analysis problem is usually understood with reference to causal inference, it is equally problematic when the objective of the research is descriptive. For example, in addressing the question of global inequality the issue of theoretical and substantive import concerns individuals. Yet data for individuals prior to the 1980s is scarce throughout the developing world. Thus, analysts are in the position of trying to infer the income status of individuals from aggregate, national-level data (GDP) – the problem of ecological inference noted above.

Cumulation

Science is not a solitary venture; it is better conceptualized as a collaborative project among researchers working on a particular subject area. This means that a research design's utility is partly a product of its methodological fit with extant work. Three elements facilitate cumulation: the *standardization* of procedures across studies; the *replication* of results; and the *transparency* of procedures.

Standardization

One of the chief avenues to collaboration is the standardization of procedures across research designs. If there is a usual way of investigating a particular issue this should be slavishly imitated, at least as a point of departure, for the standardization of approaches provides a benchmark against which new findings can be judged.

This may sound like a recommendation for theoretically modest exercises that merely re-test old ideas. It is not. Recall that in this section we are discussing criteria relevant to theory appraisal, not theory construction. We assume that a theory (and a more specific hypothesis or set of hypotheses) is already at hand. Given this theory – be it bold and original, or tamely derivative – it is advisable to standardize the research design as much as possible, at least at the outset.

The standardization of research designs allows findings from diverse studies to cumulate. Consider that if each new piece of research on vouchers utilizes idiosyncratic input and output measures, background controls, and other research design features, our knowledge of this topic is unlikely to move forward. A thousand studies of the same subject – no matter how impeccable their internal validity – will make only a small contribution to the growth of knowledge about vouchers if they are designed in *ad hoc* (and hence incommensurable) ways.

Novelties must be distinguishable from original contributions, and the question is assessable only insofar as a study can be measured by the yardsticks provided by extant work on a subject.

The call for standardization is a call for a more organized approach to knowledge-gathering. Richard Berk notes the great potential gains that might be realized from "suites of studies carefully designed so that variants in the interventions [can] be tested with different mixes of subjects, in different settings, and with related outcomes, all selected to document useful generalization targets."[21] So constructed, the possibilities for meta-analysis are vastly enhanced, and with it the prospect of theoretical advance.

Unfortunately, in the current highly individualized world of social research it is virtually impossible to aggregate results emanating from separate studies of the same general subject, for each study tends to adopt an idiosyncratic set of procedures.[22] In contrast to the natural sciences, there appears to be very little premium on standardization in the social sciences. Yet the case for standardization seems strong. Just as theories should fit within a broader theoretical framework – the criterion of *commensurability*, discussed in Chapter 3 – research designs should fit within the broader framework within which a particular issue has been addressed.

Replication

Another way that scientific activity relates to a community of scholars is through the replication of results. This project of replication takes place at two stages: (a) at the beginning of a study, as a way to verify extant findings in a new venue; and (b) after a study has been completed, as a way of testing that study's internal and external validity. (If replication is conducted during a study it is likely to be referred to as robustness testing, discussed in Chapter 10.[23])

Research on a topic typically begins by replicating key findings related to that research. To be sure, not all subjects have "findings" in the natural-science sense. Yet most fields recognize a set of propositions that are widely believed to be true; we shall call them findings even if they are closer to common-sense beliefs. Whatever the terminology, it is helpful if new research on a topic

[21] Berk (2005: 16). See also Berk *et al.* (1992); Bloom, Hill, and Riccio (2002).

[22] Briggs (2005); Petitti (1993); Wachter (1988). One possible exception to this pessimistic conclusion may be found in the field of experimental studies that have been conducted over the past few decades on subjects such as voter turnout (see the GOTV web site maintained by Don Green at Yale: http://research.yale.edu/GOTV) or employment discrimination (Pager 2007).

[23] Firebaugh (2008: ch. 4).

begins by exploring these well-known hypotheses. Are they true *here* (in this setting)? This will help clarify the validity of the chosen research design, not to mention the validity of the previous finding. This is the initial replication.

Other replications occur after a study has been completed, either prior to or after publication. (This is the more usual employment of the term.[24]) In order to facilitate replication, a research design must be conducted in such a way that future scholars can reproduce its results. Consider that findings are likely to remain suspect until they can be replicated – perhaps multiple times. We are cognizant that any number of factors might have interfered with the validity of any particular study, including (among other things) measurement error and the willful mis-reporting of data. Verification involves repetition; claims to truth, therefore, involve assurances of replicability. If a finding is obtained under circumstances that are essentially un-repeatable, then we rightfully entertain doubts about its veracity. This conforms to the narrow understanding of replication – the ability of future researchers to replicate a study's findings by carefully following the methods of procedure and sources of data that were originally employed.

But replication does not refer only to the narrowly circumscribed reiteration of a study, in near-identical circumstances. It also refers to the *variations* that may be – and ought to be – introduced to the original study. Paul Rosenbaum comments:

> The mere reappearance of an association between treatment and response does not convince us that the association is causal – whatever produced the association before has produced it again. It is the tenacity of the association – its ability to resist determined challenges – that is ultimately convincing.[25]

A finding that persists in the face of dramatic alterations in setting (background conditions), measurement instruments, specification, and treatment strength is a finding that is strongly corroborated. It is much more likely to be true than a finding that has been replicated in only minor respects. In this vein, it is important to note that replications offer not only a way to check a study's internal validity but also a means of testing – and where necessary, re-evaluating – a study's external validity. What are the boundaries of a theory?

Granted, some styles of research are easier to replicate than others. Experiments and large-*N* observational studies are replicable to a degree that qualitative work is generally not. However, in the case of large-*N* observational studies the meaning of "replication" is usually understood in a fairly restrictive fashion, that is, taking

[24] Freese (2007); King (1995); King, Keohane, and Verba (1994: 23, 26, 51).

[25] Rosenbaum (2010: 103).

the author's dataset (or a similar dataset) and replicating the author's results. This is a fairly mechanical procedure. For example, in replicating a cross-national statistical study of economic development and democracy a scholar might try to replicate extant findings and then proceed to make small alterations – adding countries (with imputed data), adding years, or using different measures of democracy.

By contrast, the replication of qualitative work is usually understood to involve the data-collection phase of research, which may be archival, ethnographic, or discursive. For example, a serious attempt to replicate James Mahoney's historical work on democratization in Central America would presumably involve a review of the author's extensive list of primary and secondary sources, and perhaps additional sources as well.[26] This represents months of research, and is not at all mechanical.[27]

The equivalent data-gathering replication in a large-*N* setting would be to re-code all the data for a key variable. In our previous example this might mean re-coding the democracy variable for all countries and all years. This is not what is usually intended by replication in a quantitative context. But there is no reason not to apply the concept of replication to this commendable cross-checking of findings.

Whatever the difficulties and ambiguities, replicability is an ideal for which all research ought to strive. Arguably, it is even more important for qualitative work than for quantitative work, given the degree of authorial intervention that is usually involved in the latter (and hence the greater possibility of investigator bias). Historical researchers should include scrupulous and detailed footnotes of their sources so that future scholars can re-trace their steps. Interview-based work should include notations about informants so that future researchers can locate these people. They may also put on file their set of notes, transcripts (or recordings) of interviews – whatever might be useful for purposes of replication (without compromising the identities of sources whose secrecy has been promised).[28]

Transparency

Evidently, standardization and replication are possible only insofar as procedures employed in empirical analyses are transparent to scholars. One cannot

[26] Mahoney (2002).

[27] An example of this sort of replication can be found in Lieshout, Segers, and van der Vleuten (2004), an attempt to replicate the archival work of Moravcsik (1998).

[28] See Hammersley (1997); Mauthner, Parry, and Backett-Milburn (1998), and the articles in Corti, Witzel, and Bishop (2005).

standardize or replicate what is ambiguous. Thus, implicit in the call for cumulation is the call for *transparency*. "The pathway between the data and the conclusions should be . . . clear."[29] For, without transparency, no finding can be fully evaluated.

It is common in natural sciences for researchers to maintain a laboratory notebook in which a close record is kept of how an empirical analysis unfolds. While it may not be necessary to record every specification test, it should at least be possible for future scholars to see which tests were conducted, in what order, and with what implications for the theory. By contrast, if scholars see only the final product of a piece of research (which may have unfolded over many years) it is more difficult to render judgment on its truth-value. One fears, in particular, that the final data tables may contain the one set of tests that culminated in "positive" (i.e., theoretically significant) results, ignoring hundreds of prior tests in which the null hypothesis could not be rejected.

Granted, the achievement of full transparency imposes costs on researchers, mostly in the form of time and effort (since the posting of notebooks is essentially cost-less). And it does not entirely solve problems of accountability. Someone must read the protocols, an investment of time. Even then, we shall never know if all procedures and results were faithfully recorded. However, the institution of a transparency regime is a precondition of greater accountability, and may in time enhance the validity and precision of empirical analysis in the social sciences.

Theoretical fit

Recall that the purpose of an empirical analysis is to shed light on an argument or theory. The relationship of the test to the argument is, therefore, a particularly sensitive issue. Three issues bear on the theoretical fit of a research design: *construct validity*, *severity*, and *partition*. All may be considered aspects of a general scientific ideal known as the *crucial* (or critical) *test*.[30]

Construct validity

Construct validity refers to the faithfulness of a research design to the theory that is under investigation.[31] This includes concept validity: the operationalization of

[29] Cox (2007: 2), quoted in Rosenbaum (2010: 147).

[30] Eckstein (1975); Forsyth (1976); Popper (1965: 112). Platt (1964) suggests that the notion may be traced back to Francis Bacon.

[31] Shadish, Cook, and Campbell (2002).

a key concept with a set of indicators. But it also includes basic assumptions or interpretations of the theory. Consider that if a research design deviates significantly from the theory – involving, let us say, questionable assumptions about the theory or building on peripheral elements of the theory – then the theory can scarcely be proven or disproven, for the research design does not bear centrally upon it. By the same token, if a researcher chooses a hypothesis that lies at the core of a theory, the research design has greater relevance.

In this context, one might contemplate the vast range of work on education policy that bears in some way or another on vouchers.[32] A good deal of this research lies at the periphery of the core hypothesis about school vouchers and school performance; it is somewhat relevant, but not primary. For example, if a study shows that vouchers have no effect on racial harmony in schools this finding, while interesting, is not likely to be considered central to the theory. As such, the theory is relatively unaffected by the finding. If, by contrast, a study shows that vouchers have no effect on educational performance this is devastating to the theory, precisely because the research design and the theory are so closely aligned.

Granted, many grand theories do not rest on a single central hypothesis (such as vouchers and educational performance). Consider the larger theory of free market competition that informs the voucher idea. This theory, as framed by Milton Friedman, Friedrich von Hayek, or Adam Smith, is not amenable to any knock-down tests of which I am aware. Capitalism, like socialism, resists falsification. Evidently, the more abstract the theory, the harder it is to translate that theory into a viable empirical test.[33] Even so, researchers must work hard to ensure that empirical tests are not theoretically trivial. A high level of internal and external validity will not rescue a theoretically irrelevant study, for which we reserve the epithet "straw-man."

Severity

Some empirical tests are easy, requiring little of a theory to clear the hurdle (which may or may not be formalized in a statistical test such as a *t*-test). Other empirical tests are hard, requiring a great deal of a theory. *Ceteris paribus*, we are more likely to believe that a theory is true when it has passed a severe empirical test (as long as the test has some degree of construct validity). "Confirmations should count," insists Popper,

[32] Daniels (2005).

[33] Gorski (2004); Green and Shapiro (1994); Lieberson (1992).

> only if they are the result of *risky predictions*; that is, if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory – an event which would have refuted the theory.[34]

The same factors work in reverse if one is attempting to disprove (falsify) a theory. If the theory fails a very hard test, one may not be inclined to conclude that it is wrong. If, on the other hand, it fails an easy test – one that, according to the premises of the theory it ought to have passed – then one's attitude toward the theory is apt to be more skeptical.

An analogy drawn from track-and-field may help to illustrate the point. Suppose, for example, we wish to test the relative ability of various athletes in the high jump, an event that traces its lineage to ancient Greece. In the first test, we set the bar at 10 ft (3 m) – a ridiculous goal, given that the highest recorded free jump is just over 8 ft (2.5 m). Predictably, all the athletes fail to clear this most-difficult test. In the second test, we approach the matter differently, setting the bar at 3 ft (1 m). Predictably, all the athletes clear this least-difficult test. Evidently, we have learned nothing whatsoever of the relative abilities of this group of athletes at the end of these two tests. To be sure, had any of these athletes passed the hard test (or failed the easy test) we would have learned, beyond a shadow of a doubt, that that particular athlete was an extraordinarily good (bad) high jumper. This is the irony of the criterion of *severity*: it depends on the outcome of the test. Otherwise stated, one wishes to set the bar just high enough that it can be cleared by some people (but no higher), or just low enough that it cannot be cleared by some people (but no lower).

One apparent resolution of this problem is to avoid setting arbitrary thresholds. Instead, ask athletes to jump as high as they can and simply measure their relative performance – a *continuous* metric. Or, if circumstances demand (e.g., if it is necessary to establish a bar in order to measure the height of a jump), set up numerous tests with varying thresholds. These two approaches amount to the same thing, except that the latter requires multiple iterations and is in this sense less efficient.

A flexible approach to testing is justified in many contexts. However, the sacrifice one makes in adopting a flexible standard should be clear. Wherever the criteria for success and failure are not spelled out clearly in advance the resulting research is less falsifiable, that is, more liable to varying interpretations of success and failure.

[34] Popper (1965: 36). See also Popper ([1934] 1968); Howson and Urbach (1989: 86); Mayo (1996: ch. 6); Mayo and Spanos (2006).

Moreover, even if one eliminates an *a priori* threshold for success/failure, many factors are likely to remain that serve to structure the degree of difficulty of a test. Returning to our track-and-field example, it will be seen that athletes' performance is affected by a great many "contextual" factors – altitude, whether the event is held indoors or outdoors, the quality of the surface, the audience in attendance, and so forth. Relative performance varies with all of these factors (and perhaps many more). In social science settings, the list of contextual factors is also quite large. Here one might consider various research design factors that "load the dice" for, or against, a school vouchers study. Suppose, for example, that a study of vouchers is conducted in a community where teachers and administrators, as well as many of the participants in the program, are skeptical about – and even downright hostile to – the reform. Or suppose that teachers working in vouchers schools (schools attended by children with vouchers) are less experienced or less educated than teachers working in public schools. Suppose, finally, that the monetary value of the voucher that students received was minimal – less than prior work and theory suggests would be necessary to achieve significant changes in student achievement. These are all factors that would seem to load the dice against a positive finding. If, under the circumstances, that study finds that vouchers induce a positive (and statistically significant) effect on student performance, we are likely to be especially impressed by the finding. On the other hand, if the foregoing factors are reversed, and the bias of a study appears to *favor* the vouchers hypothesis, a positive finding will have little credibility. Indeed, it is quite likely spurious.

Assumptions about the direction of probable bias may play an important role in evaluating the empirical findings of a study (*ex post*), as well as in designing a study (*ex ante*). Rosenbaum notes that a

> sometimes compelling study design exploits a claim to know that the most plausible bias runs counter to the claimed effects of the treatment. In this design, two groups are compared that are known to be incomparable, but incomparable in a direction that would tend to mask an actual effect rather than create a spurious one. The logic behind this design is valid: if the bias runs counter to the anticipated effect, and the bias is ignored, inferences about the effect will be conservative, so the bias will not lead to spurious rejection of no effect in favor of the anticipated effect.[35]

In short, the degree of difficulty imposed by a research design with respect to a particular hypothesis is an intrinsic part of any study. Whether the purpose of

[35] Rosenbaum (2010: 123).

the research is positive (to prove a causal proposition) or negative (to disprove a causal proposition), the value of a research design derives partly from its relative "crucial-ness." The following question thus arises with respect to any study: how likely is it that theory A is true (false), given the evidence? The harder (easier) the test, the more inclined we are to accept the conclusion – *if* the test is passed (failed).

Even if one dispenses with arbitrary thresholds for judging success and failure, it will still be the case that background factors built into a research design qualify that test as "easy" or "difficult" with respect to a particular hypothesis. These factors, which move well beyond the narrow issues addressed by quantitative measures of statistical significance or statistical power, must be taken into account if we are to arrive at a judgment of the overall truth-value of a finding. Such issues beg consideration *ex ante*, during the design of a study, and *ex post*, as researchers assess a study's contribution.

Whether one opts for a research design that leans toward greater or lesser difficulty depends upon many factors. Easy tests are often appropriate at early phases of hypothesis testing, when a project is still largely exploratory and when few extant studies of a subject exist. Hard tests become appropriate as a hypothesis becomes well established and as the number of extant studies multiplies.

Of course, hard tests are better if they can be devised in a way that is fair to the theory under investigation – if they maintain *construct validity*, in other words. A good deal of research in the natural sciences seems to follow this model. Consider this list of risky predictions that served to confirm or refute important theories in physics:

> Newton's prediction of elliptical orbits of the planets from the inverse square law of gravitation; various experiments confirming the wave theory of light; Maxwell's prediction of electromagnetic waves from a mathematical model; the Michelson–Morley experiment that disproved the existence of the ether and confirmed the constant velocity of light; Kelvin's prediction of absolute zero temperature; derivations from Poisson's and Fourier's mathematical theory of heat; inferences based on the kinetic theory of gases and statistical mechanics; the prediction of various subatomic particles; Gamow's prediction that the Big Bang had left its mark in radiation at the edge of the universe; and, most famously, Einstein's predictions that led to the confirmation of his special and general theories of relativity, such as the "bending" of a star's light by gravitational attraction.[36]

[36] Coleman (2007: 129–130).

The author of this compendium, Stephen Coleman, also helpfully identifies several features of these theoretical predictions that proved useful in establishing a crucial test. These include:

• Prediction of a constant or invariant (like the speed of light or a freezing point) • Prediction of a specific number • Prediction of a symmetry, often derived from a mathematical model • Prediction of a topological fixed point • Prediction of a limit or constant, or dynamic limit cycle • Prediction of a specific or unusual dynamic behavior pattern • Prediction of a specific spatial (geographic) pattern • Prediction of a statistical distribution, possibly an unusual distribution • Prediction that data will have a "signature" – a unique mathematical shape (as used for detecting heart arrhythmias, nuclear tests, tsunamis, or submarines).[37]

These are useful exemplars and suggestions. It is especially important to appreciate that there are a multitude of ways to construct a test for a given hypothesis, only one of which takes the form of a classic linear and additive model. A common approach is to specify (or examine for clues, *ex post*) a dose–response relationship, that is, the way in which *Y* responds to a change in *X*.[38] Many of these alternatives offer a higher degree of falsifiability because they offer highly specific predictions, drawn directly from the theory – predictions that are unlikely to be true unless the theory is true – as opposed to the run-of-the-mill social science prediction that "an increase in *X* will lead to an increase in *Y*."

Of course, one may be skeptical about the practicality of this advice.[39] How many social phenomena are amenable to precise *a priori* predictions? How many are amenable to mathematical models of the sort that would yield precise, *a priori* predictions? The present state of formal modeling in most social science disciplines, while aiming to achieve the crucial tests of physics, is still a long way from that goal.

We do not need to resolve this question. For present purposes, it is sufficient to observe that the precision of a theory is essential to the severity of a test. Both are a matter of degrees, and both are a key component of that theory's falsifiability.

Partition

Falsifiability is also enhanced insofar as an argument can be effectively isolated, or *partitioned*, from the empirical analysis. This reduces the possibility that a theory might be adjusted, *post hoc*, so as to accommodate negative

[37] Coleman (2007: 130). See also Taagepera (2008).

[38] Rosenbaum (2010: 124–125).

[39] Grofman (2007).

findings. It also reduces the temptation to construct arguments closely modeled on a particular empirical setting ("curve-fitting"), or research designs whose purpose is to prove (rather than test) a given argument. Ideally – at least for purposes of appraisal – the construction of an argument should be considered a separate step from the testing of that same argument.[40]

Another sort of partition can sometimes be erected between the research design phase of a study and the data analysis phase of a study. This distinction – between prospective design and retrospective analysis – is a hallmark of the experimental method, and one of the reasons why experiments are rightly regarded as enhancing the falsifiability of a study.[41] There is less opportunity for *ex post facto* adjustments of design to rectify inconvenient empirical results.

Granted, the goal of partitioning is always a matter of degree. It is not clear how the advance of knowledge could occur if partitions were to be complete and final. (What does "final" mean?) Note that any failed test (not to mention successful tests) must be followed up with further tests, and these further tests must take the failures (and successes) of the past into account. In this sense, all research is an iterative process, moving back and forth between theory and evidence.

The criterion of partition may be understood, first, as referring to the length of time that ensues between initial testing and subsequent reformulation and re-testing. If the duration is minute – for example, statistical specification tests conducted at intervals of several seconds through an automated routine – then we are apt to label the procedure curve-fitting. One is not really testing a model; one is finding the best fit between a set of variables (representing a set of very loose hypotheses) and a sample of data. If, on the other hand, the duration is lengthy – say, a year or more – then we would be more inclined to feel that the goal of partition has been achieved. Theory formation has been segregated from theory-testing.

Second, partition refers to data employed for testing. Ideally, arguments should be tested with a sample of observations different from those employed to generate the theory. This provides *out-of-sample* tests. To be sure, if samples are large and representative this should not make much difference; the same results should obtain. And if samples are small and/or non-representative,

[40] King, Keohane, and Verba (1994) advise: "Ad hoc adjustments in a theory that does not fit existing data must be used rarely" (p. 21). "Always . . . avoid using the same data to evaluate the theory [you] used to develop it" (p. 46). Original data can be reused "as long as the implication does not 'come out of' the data but is a hypothesis independently suggested by the theory or a different data set" (p. 30). See also Eckstein (1992: 266); Friedman ([1953] 1984: 213); Goldthorpe (1997: 15).

[41] Rubin (2008: 816).

a strong argument can be made for combining all available data into a single sample – thereby maximizing sample size and representativeness. So, one may be skeptical of how practical the out-of-sample test is in practice. Nonetheless, where practicable, it is certainly desirable.

Finally, and most importantly I think, partition refers to a state of mind. Insofar as theorizing and testing are separable, the most important feature of this separation is not the length of time that one is segregated from the other or the difference in samples, but rather the attitude of the researcher.

Mental partition requires multiple personalities. At the stage of theory-generation, the researcher must be nurturing – a booster of the theory that is being created. All efforts are focused single-mindedly on the creation and sustenance of that new and still fragile idea. *A priori* speculations about the world are *de rigueur*, for one must posit a great deal in order to establish the foundation for a theory. Arguments are argumentative.

At the stage of theory-testing, by contrast, a second personality must be adopted. This personality is non-partisan, or perhaps even openly skeptical with respect to the main hypothesis under examination. The baby has been born, it has suckled, it is now strong enough to face the rigors of the world (i.e., empirical testing). To continue the metaphor, good research requires killing one's own children from time to time.

This is the sort of mental partition that research requires. Arguably, it is only fully achievable when the two stages of research – theory-formation and theory-testing – are carried out by different persons, that is, where the tester has no incentive to disprove the null hypothesis. But in the real world of research, especially social science research (where funding and personnel are limited relative to the number of research questions under consideration), this is rarely possible. So, we must appeal to the researcher's good sense and to his or her capacity to transition from the mentality of theorizing and nurturing to the mentality of analysis and severe tests, that is, from discovery to appraisal (Chapter 2).

It is vital that the audience for a piece of research feel confident in the impartiality of the researcher throughout the testing phase. There are many ways in which researcher bias can creep in, and there is no way for audiences to monitor the situation if researchers are in charge of testing their own hypotheses. Principal–agency complications are too great. This means that trust is required, and the researcher must work hard to earn the audience's trust.

One technique is to declare one's biases at the outset, so that it is clear to the reader of a report where the researcher's point of departure is (and so that the distinction between theorizing and testing is preserved, at least rhetorically). If it happens that a research finding runs *counter* to the original hypothesis, audiences may be more inclined to believe that result, on the assumption that it has cleared an especially high hurdle (or, at the very least, that investigator bias has not infected the result). In situations of poor oversight, the mind-set of the researcher is highly relevant to an *ex post* analysis of findings.

Part II

Description

5 Concepts

The history of the social sciences is and remains a continuous process passing from the attempt to order reality analytically through the construction of concepts – the dissolution of the analytical constructs so constructed through the expansion and shift of the scientific horizon – and the reformulation anew of concepts on the foundations thus transformed . . . The greatest advances in the sphere of the social sciences are substantively tied up with the shift in practical cultural problems and take the guise of a critique of concept-construction.

Max Weber[1]

As we are . . . prisoners of the words we pick, we had better pick them well.

Giovanni Sartori[2]

Description will be understood in this book as any empirical argument (hypothesis, theory, etc.) about the world that claims to answer a *what* question (e.g., *how*, *when*, *whom*, or *in what manner*). By contrast, wherever there is an implicit or explicit claim that a factor generates variation in an outcome the argument will be regarded as causal. The distinction between these two key concepts thus hinges on the nature of the truth-claim – not on the quality of the evidence at hand, which may be strong or weak.[3] Description

[1] Weber ([1905] 1949: 105–106).

[2] Sartori (1984: 60).

[3] This is somewhat at variance with current linguistic practices, where these terms are frequently employed as a signal of the quality of the evidence at hand: with "causal" reserved for experimental or quasi-experimental evidence and "descriptive" reserved for evidence that is (for whatever reason) weak. Andrew Gelman advises: "When describing comparisons and regressions, try to avoid 'effect' and other causal terms (except in clearly causal scenarios) and instead write or speak in descriptive terms": www.stat.columbia.edu/~cook/movabletype/archives/2009/03/describing_desc.html. In this vein, some researchers prefer to regard *all* evidence as descriptive, so as to emphasize the interpretive leap that causal inference requires (Achen 1982: 77–78). The evident problem with this definitional move is that it deprives us of a way of distinguishing between arguments that embrace different goals. Note that any attempt to appraise the truth-value of an empirical proposition must begin by resolving what the goals of that proposition are, i.e., descriptive, causal, or some other. If the truth-claim is unclear then it is impossible to falsify. From this perspective, preserving the traditional distinction between *what* questions and *why* questions ought to be a high priority for the discipline.

is the topic of Part II, while causation is the topic of Part III. Description rightly comes first; one must describe in order to explain (causally). However, the reader will find many comparisons and contrasts across the two topics interwoven throughout the book.

Because this book is focused on generalizing statements about the world (Chapter 1), I am not concerned with descriptions that reflect only on individual cases or events (without any attempt to exemplify larger patterns).[4] Consequently, in this book description is always an *inferential* act. To generalize is to infer from what we know (or think we know) to what we do not know.[5] One sort of inferential leap is from observations within a sample that are deemed secure to those that are uncertain or missing (problems of "measurement error" or "missing data") and to dimensions that are inherently unobservable ("latent characteristics"). Another sort of inferential leap is from a studied case or sample to a larger (unstudied) population. In both respects, descriptive models offer a "theory" about the world,[6] "a 'formula' through which the data can be reproduced."[7]

In recent years, the quest for scientific understanding has come to be equated with the quest for a causal understanding of the world across the social sciences. By contrast, the task of description is identified with idiographic storytelling – impressionistic narratives relating details about particular times and places – or with issues of measurement. The term itself has come to be employed as a euphemism for a failed, or not yet proven, causal inference. Studies that do not engage causal or predictive questions are judged "merely" descriptive.[8] Likewise, evidence for a causal proposition that is judged especially weak is likely to be characterized as "descriptive." More generally, the view of description that obtains in the social sciences (and especially in economics and political science) is of a mundane task – necessary, to be sure, but of little intrinsic scientific value.

The subordination of description to causation is problematic from a number of perspectives. First and foremost, a large class of descriptive topics is

[4] To reiterate: this does not preclude the discussion of particular events and outcomes, but it does mean that the goal of these cases is to reflect upon the characteristics of a larger population.

[5] On some fundamental level, all empirical knowledge may be considered inferential. However, it is helpful to distinguish between readily apprehensible facts about the world ("observables") and those which must be speculated upon ("unobservables"). I reserve the concept of inference for the latter.

[6] Jacoby (1999).

[7] Berk (2004: 207).

[8] It is not clear when, precisely, this pejorative connotation arose. It was invoked, or commented on, in the social science literature at various points in the mid- to late twentieth century (e.g., Klimm 1959; Sen 1980; Singer 1961). However, it probably stretches back further in time within the tradition of Anglo-American economics and political science (e.g., Clark and Banks 1793: 157).

intrinsically important. Into this class fall subjects like democracy, human rights, war, revolution, standards of living, mortality, ethnic conflict, happiness/utility, and inequality. These topics (and many others) deserve to be explored descriptively. We need to know how much democracy there is in the world, how this quantity – or bundle of attributes – varies from country to country, region to region, and through time. This is important regardless of what causes democracy or what causal effects democracy has.[9]

The concern is that if conceptualization and measurement of democracy occurs only in the quest for causal inference we may not achieve the same level of accuracy, precision, and comprehensiveness with respect to the topic. A research agenda motivated solely by a causal hypothesis is apt to take short-cuts when it comes to describing the left- and right-hand variables. Moreover, that which one chooses to describe may be influenced by the general *X*/*Y* relationship one expects to find, and this may introduce biases into how we describe the phenomenon. To be sure, there is nothing wrong with causally oriented description. But it may pose a problem if this is the principal means of approaching a topic within a field over many years.[10]

A second reason for liberating description from specific causal hypotheses is practical in nature. Often, it is more efficient to collect evidence when the objective of the investigation is descriptive rather than causal. Consider that

[9] For examples of natural science research that is descriptive rather than causal see Bunge (1979).

[10] Naturally, if the social sciences were grounded in a single causal-theoretical framework on the order of evolution within the biological sciences then we would possess a causal model around which a coherent description of the world might be reliably constructed. However, we lack such a unifying paradigm, and in its absence it is difficult to say how a causally ordered description of the political world might be organized or what it would look like (in concrete terms). One might counter that in a multiparadigmatic universe one should look to smaller-scale causal hypotheses to organize the work of the discipline, along the "behavioralist" model. But here one stumbles upon another problem of indeterminacy. Because causal attribution is difficult to establish for most nontrivial questions in social science it is problematic to assert that *X* matters as a subject of investigation only insofar as it causes *Y* (or *Y* matters only insofar as it is caused by *X*). Ambiguity about whether *X really* causes *Y* means that it may be safer to approach *X* and *Y* first as descriptive phenomena – important in their own right – rather than as potential independent and dependent variables. As an example, let us reconsider the question of "democracy." Presumably, this feature has many causal properties. However, we do not know for sure what these are; and certainly, we do not know *precisely* what they are. Consequently, the subject is perhaps better approached, at least initially, as a descriptive issue. Of course, I do not mean to suggest that descriptive inference be carried out in ignorance of all causal potentialities. I mean, rather, that in circumstances where causal frameworks are open-ended – presumably the vast majority of cases in social science – descriptive inference ought to be carried out independent of any *particular* causal hypothesis. This helps to avoid a highly prejudiced (i.e., particularistic, idiosyncratic) definition of a subject matter. All plausible causal hypotheses are relevant – those in which a subject serves as an independent variable, those in which it serves as a dependent variable, and those in which it serves as a causal pathway in some larger subject. When considered in this open-ended fashion the subject of interest (e.g., democracy) is rightly approached descriptively rather than simply as an adjunct to subsequent causal analysis.

data is collected from persons, governments, archives, and other organizations. Collecting evidence from these sources in a systematic fashion requires considerable energy and resources, sustained over many years. When a data-collection effort is constructed around a single causal hypothesis or theory the scholar's purview is naturally quite limited; only those factors having direct bearing on the hypothesis will be collected. This may be efficient in the short run, but it is not likely to be efficient in the long run. Narrowly focused data expeditions entail scaling high cliffs and returning to base camp with only a small sample of what one finds at the peak. Later expeditions, focused on different hypotheses, will require re-scaling the same peak, a time-consuming and wasteful enterprise. By contrast, if an evidence-gathering mission is conceptualized as descriptive rather than causal (which is to say, no *single* causal theory guides the research), it is more likely to produce a broad range of evidence that will be applicable to a broad range of questions, both descriptive and causal.[11]

In sum, there are good reasons to approach description as a distinctive – and essential – task of social science. This is the motivation of Part II of the book. This chapter focuses on social science concepts, the linguistic containers we use to carve up the empirical world. Chapter 6 offers a typology of descriptive arguments, and Chapter 7 focuses on the task of measurement, the "analysis" of descriptive propositions.

The quandary of description

Conventional wisdom presumes that causal inference is harder, methodologically speaking. "*What* questions are generally easier to answer than *why* questions" states Glenn Firebaugh.[12] "Empirical data can tell us what is happening far more readily than they can tell us why it is happening," affirms Stanley Lieberson.[13] Reading the methodological literature, one might infer that description is a relatively simple and intuitive act of apperception.

And yet, many descriptive questions circulating through the disciplines of social science are recalcitrant. Consider the following:

(1) Do voters conceptualize politics ideologically[14] or nonideologically?[15]
(2) Is global inequality increasing[16] or remaining about the same?[17]

[11] Schedler (forthcoming).
[12] Firebaugh (2008: 3).
[13] Lieberson (1985: 219). See also Gelman (2010).
[14] Nie, Verba, and Petrocik (1976).
[15] Converse (1964).
[16] Milanovic (2005).
[17] Bourguignon and Morrisson (2002); Dollar (2005); Firebaugh (2003).

(3) Is American political culture liberal/egalitarian,[18] republican,[19] or a mixture of both, along with various ascriptive identities?[20]

These are all essentially descriptive questions about the social world (though, to be sure, they contain causal implications). They have also proven to be hotly contested. And they are not unusual in this regard. A random sample of (nontrivial) descriptive arguments would likely reveal a high level of uncertainty. Indeed, there is great consternation over the poor quality and measly quantity of evidence by which we attempt to make sense of the social world.[21] Descriptive accounts of mid-level phenomena like corruption, campaign finance, civil service protection, judicial independence, and party strength are often highly problematic, or are restricted in purview to very specific contexts (and hence resist generalization). And the big concepts of social science – such as democracy and governance – have no standard and precise meaning or measurement.[22] Meanwhile, whole tracts of social and political activity remain virtually *terra incognita*.[23] As a result, empirical phenomena on the left and right sides of the typical causal model are highly uncertain. To paraphrase Giovanni Sartori, the more we advance in causal modeling, the more we leave a vast, uncharted territory at our backs.[24]

To get a glimpse of the methodological problems we face in reaching descriptive inferences let us contrast the following two questions:

(1) What is democracy, and how might it be operationalized?
(2) Does democracy enhance the prospect of peaceful coexistence?

Note that the causal question (2) presumes an answer to the descriptive question (1). In order to estimate democracy's causal effect one must first establish the definition and measurement of this vexing concept. Logic suggests that if Proposition 2 builds on Proposition 1 it must be at least as difficult to prove as Proposition 1. And yet, by all appearances, there is greater scholarly consensus on the answer to question (2) than on the answer to question (1). Scholars of

[18] Hartz (1955); Tocqueville (1945). [19] Pocock (1975). [20] Smith (1993).

[21] Heath and Martin (1997); Herrera and Kapur (2007); Kurtz and Schrank (2007); Munck (2009); Rokkan *et al.* (1970: 169–180).

[22] On democracy, see Bowman, Lehoucq, and Mahoney (2005); Coppedge (forthcoming); Hadenius and Teorell (2005); Munck (2009); Munck and Verkuilen (2002). On governance, see Kurtz and Schrank (2007); March and Olson (1995); Pagden (1998); Pierre (2000). A wide-ranging compendium of indicators for democracy and governance can be found in USAID (1998).

[23] As one example one might consider local government in the developing world, a topic that has elicited little systematic empirical attention, despite its evident importance. For a recent review of this neglected field of study see UN Habitat (2004).

[24] Sartori (1970: 1033).

international relations generally agree that regime status has a causal effect on peace and war such that democracies are less likely to fight wars with one another, all other things being equal. Whether or not democracy is a *sufficient* condition for peace may never be determined, and scholars continue to debate the causal mechanisms at work in this relationship. However, there is still a large measure of agreement on the democratic peace as – at the very least – a probabilistic causal regularity.[25] All things being equal, two democratic countries are less likely to go to war with one another than two countries, one or both of which are nondemocratic. By contrast, no such consensus exists on how to conceptualize and measure democracy. The causal proposition is fairly certain, while the descriptive proposition that underlies it is highly uncertain.

This is the paradoxical pattern for many descriptive inferences. Despite the fact that causal inferences build on descriptive inferences the former are often more certain and more falsifiable. The reasons for this are partly intrinsic to the enterprise. For example, descriptions often center on matters of definition, and therefore are not as amenable to appeals to evidence. Descriptions are also often exploratory in nature, and therefore constructed in close contact with the evidence (a problem of insufficient *partition* [Chapter 4]).

That said, some of the methodological problems encountered by descriptive inference are remediable. Arguably, they are a product of the general lack of methodological self-consciousness that permeates this enterprise. My hope is that by clarifying the common criteria pertaining to descriptive arguments, and by classifying the immense variety of descriptive arguments, we may improve the quality of descriptive inference – and, perhaps, over time, enhance its standing in the social sciences.

Concepts

Concept formation lies at the heart of all social science endeavors.[26] It is impossible to conduct work without using concepts. It is impossible even to conceptualize a topic, as the term suggests, without putting a label on it. Concepts are integral to every argument for they address the most basic question of social science research: what are we talking about?

If concepts allow us to conceptualize, it follows that creative work on a subject involves some *re*conceptualizing of that subject. A study of democracy, if persuasive, is likely to alter our understanding of "democracy," at least to some

[25] Brown, Lynn-Jones, and Miller (1996); Elman (1997).

[26] Sartori (1970: 1038).

degree.[27] No use of language is semantically neutral. Authors make lexical and semantic choices as they write and thus participate, wittingly or unwittingly, in an ongoing interpretive battle. This is so because language is the toolkit with which we conduct our work, as well as the substance on which we work. Progress in the social sciences occurs through changing terms and definitions. This is how we map the changing terrain (or our changing perceptions of the terrain).

Unfortunately, all is not well in the land of concepts. It has become a standard complaint that the terminology of social science lacks the clarity and constancy of natural science lexicons. Concepts are variously employed in different fields and subfields, within different intellectual traditions, among different writers, and sometimes – most alarmingly – within a single work. Concepts are routinely stretched to cover instances that lie well outside their normal range of use.[28] Or they are scrunched to cover only a few instances – ignoring others that might profitably be housed under the same rubric. Older concepts are redefined, leaving etymological trails that confuse the unwitting reader. New words are created to refer to things that were perhaps poorly articulated through existing concepts, creating a highly complex lexical terrain (given that the old concepts continue to circulate). Words with similar meanings crowd around each other, vying for attention and stealing each other's attributes. Thus, we play musical chairs with words, in Giovanni Sartori's memorable phrase.[29]

A result of these pathologies is that studies of the same subject appear to be talking about different things, and studies of different subjects appear to be talking about the same thing. Cumulation is impeded and methodological fragmentation encouraged. Concepts seem to get in the way of clear understanding.

One solution to our seemingly endless conceptual muddle is to bypass conceptual disputes altogether, focusing on the phenomena themselves rather than the labels and definitions we attach to them. If, as Galileo observed, all definitions are arbitrary, then we might as well begin by recognizing this fact.[30] It is commonly said, for example, that one can prove practically anything simply by defining terms in a convenient way. This is what prompts some commentators to say that we ought to pay less attention to the terms we use, and more to the things out there that we are talking about. "Never let yourself be goaded into taking seriously problems about words and their meanings," Karl Popper warns. "What must be taken seriously are questions

[27] Discussion of the concept of democracy in this chapter and the next draws on Coppedge (forthcoming); Coppedge and Gerring (2011); Munck (2009).

[28] Collier and Mahon (1993); Sartori (1970).

[29] Sartori (1975: 9; see also 1984: 38, 52–53).

[30] Robinson (1954: 63).

of fact, and assertions about facts, theories, and hypotheses; the problems they solve; and the problems they raise."[31]

The empiricist perspective seems reasonable on the face of things. And yet we are unable to talk about questions of fact without getting caught up in the language that we use to describe these facts. To be sure, things exist in the world separate from the language that we use to describe them. However, we cannot talk about them unless and until we introduce linguistic symbols. Any cumulation of knowledge depends upon reaching an understanding about what to call a thing and how to define it. This militates against a blithe nominalism ("call it whatever you want").

A second approach to resolving conceptual difficulty in the social sciences suggests that concept formation is irreducibly a matter of context. There is little one can say in general about concept formation because different concepts will be appropriate for different research tasks and research venues. This hoary bit of wisdom is absolutely true – but also highly ambiguous. What does context mean, and how might it help to guide the process of concept formation? I suspect that every author has their own preferred context, which means that conceptual disputes are simply displaced from "concept" to "context." Of course, I am not arguing that the choice of terms and definitions should be insensitive to research contexts. I am, rather, raising the question of precisely *how* contexts would or should guide concept formation.

A third approach to conceptual dis-ambiguation advises us to avoid high-order concepts in preference for less abstract (more "concrete") concepts. Because most of the conceptual ambiguities of social science involve large conceptual containers, such as culture, democracy, ideology, legitimacy, power, public goods, rationality, and the state, perhaps we ought to pare down our conceptual ambitions in favor of manageable units such as deaths, votes, and purchasing power. This also seems reasonable. However, there are important tradeoffs to such a strategy (known to philosophers as *physicalism*). Most obviously, we would be limited in what we could talk about. We could discuss votes but not democracy. And although this concretized lexicon might lead to greater agreement among social scientists one would have to wonder about the overall utility of a social science reconstructed along such lines. Does the act of voting matter outside a framework of democracy? Is it meaningful at all? Arguably, a social science limited to directly observable entities would have very little of importance to say. Moreover, it would have no way of putting these small-order ideas together into a coherent whole. Large-order concepts comprise

[31] Popper (1976: 19; quoted in Collier 1998).

the scaffolding on which we hang observables. Without general concepts, science cannot generalize, and without the ability to generalize science cannot theorize.[32] A social science composed purely of concrete concepts would be a series of disconnected facts and micromechanisms.

A final approach to concept dis-ambiguation seeks a taxonomic reconstruction of scientific concepts, an approach sometimes designated as "Classical" after the work of Aristotle and latter-day logicians in the Aristotelian tradition.[33] This is an attractive ideal, as the taxonomy possesses many desirable qualities (reviewed in the previous chapter). Yet while it may be practicable in some areas of natural science such as biology, the taxonomic approach does not seem to apply across the board in social science. Taxonomies have their uses, but these uses tend to be restricted to specialized settings: individual studies or very specific terrains. It is a specialized tool, not a general-purpose tool.

The general employment of social science concepts cannot be successfully contained within a set of taxonomies – much less, within a single all-embracing taxonomy. Meanings overflow the neat and tidy borders of social science taxonomies; rarely are concepts reducible to necessary and sufficient attributes. And even if social scientists were to accept such a reconstruction, one might wonder about the utility of a rigidly taxonomic lexicon. Note that the world of decisional behavior that the social sciences seek to describe and explain is characterized by a great deal of messiness and in-discreteness. Phenomena of this nature do not readily group together in bundles with clear borders and hierarchical interrelationships. Thus, while it is true that a simplified taxonomic language would reduce semantic confusion it might also reduce our capacity to correctly understand the social world. We could agree on a lot (if we all agreed to use symbols in the same way), but we could not say very much.

In this chapter I offer a somewhat new approach to the task of conceptualization. The chapter begins with a discussion of several key criteria pertaining

[32] By "theorize," I mean the search for descriptive or causal inferences that are general in scope – not the development of a theory about a single event or context. For further discussion, see Chapter 4.

[33] The classical approach to concept formation is usually traced back to Aristotle and the scholastic philosophers of the Middle Ages. Nineteenth-century exponents include Mill ([1843] 1872: 73) and Jevons (see discussion in Kaplan 1964: 68). In the twentieth century, see Chapin (1939); Cohen and Nagel (1934); DiRenzo (1966); Dumont and Wilson (1967); Hempel (1952, 1963, 1965, 1966); Landau (1972); Lasswell and Kaplan (1950); Lazarsfeld (1966); Meehan (1971); Stinchcombe (1968, 1978); Zannoni (1978); and, most importantly, Sartori (1970, 1984). For a somewhat different reconstructive approach based on the analytic philosophic tradition see Oppenheim (1961, 1975, 1981). For further discussion of the classical concept and its limitations see Adcock (2005); Collier and Levitsky (1997); Collier and Gerring (2009); Collier and Mahon (1993); Goertz (2006); Kaplan (1964: 68); Lakoff (1987); Taylor (1995).

to all empirical concepts. It continues by offering a set of strategies that may help to structure the task of concept formation in social science settings.

Criteria of conceptualization

Four elements of an empirical concept are conventionally distinguished: (a) the *term* (a linguistic label comprising one or a few words); (b) *attributes* that define those phenomena (the definition, intension, connotation, or properties of a concept); (c) *indicators* that help to locate the concept in empirical space (the measurement or operationalization of a concept); and (d) *phenomena* to be defined (the referents, extension, or denotation of a concept).

As an example, let us consider the concept of democracy. The term is "democracy." A commonly cited attribute is "contested elections." An indicator might be "a country that has recently held a contested election." And the phenomena of interest are, of course, the entities out there in the world that correspond to the concept, so defined.

When a concept is formulated (or reformulated) it means that one or all of the features is adjusted. Note that they are so interwoven that it would be difficult to change one feature without changing another. The process of concept formation is therefore one of mutual adjustment. To achieve a higher degree of conceptual adequacy one may (a) choose a different term, (b) alter the defining attributes contained in the intension, (c) adjust the indicators by which the concept is operationalized, or (d) redraw the phenomenal boundaries of the extension.

It follows that a change in any one aspect of a concept is likely to affect the other three.[34] And for this reason, our topic must be viewed holistically. It is difficult to separate out tasks that pertain only to the phenomenal realm from those that pertain to the linguistic/semantic or theoretical realms. Social science, from this perspective, is an attempt to mediate between the world of language (the term and its attributes) and the world of things (beyond language). Neither is temporally or causally prior; both are already present in a concept.

With this understanding of our task, seven criteria may be deemed critical to the formation of empirical concepts in the social sciences: (1) *resonance*, (2) *domain*, (3) *consistency*, (4) *fecundity*, (5) *differentiation*, (6) *causal utility*, and (7) *operationalization* (i.e., measurement). The last criterion forms the topic of Chapter 7, so this chapter will cover only the first six criteria. For convenience, all seven desiderata are summarized in Table 5.1.

[34] Hoy (1982).

Table 5.1 Criteria of conceptualization

1. **Resonance** (familiarity, normal usage; *antonyms:* idiosyncrasy, neologism, stipulation)
   How faithful is the concept to extant definitions and established usage?
2. **Domain** (scope)
   How clear and logical is (a) the language community(ies) and (b) the empirical terrain that a concept embraces?
3. **Consistency** (*antonym:* slippage)
   Is the meaning of a concept consistent throughout a work?
4. **Fecundity** (coherence, depth, essence, fruitfulness, natural kinds, power, real, richness, thickness)
   How many attributes do referents of a concept share?
5. **Differentiation** (context, contrast-space, perspective, reference point, semantic field)
   How differentiated is a concept from neighboring concepts? What is the contrast-space against which a concept defines itself?
6. **Causal utility** (empirical utility, theoretical utility)
   What utility does a concept have within a causal theory and research design?
7. **Operationalization** (measurement)
   How do we know it (the concept) when we see it? Can a concept be measured easily and unproblematically, i.e., without bias? (Chapter 7)

Resonance

The degree to which a term or definition makes sense, or is intuitively clear, depends crucially on the degree to which it conforms or clashes with established usage. A term defined in a highly idiosyncratic way is unlikely to be understood. At the limit – that is, with nonsense words – it is not understood at all. The achievement of communication therefore involves a search for *resonance* with established usage.[35]

Anyone inclined to discount the importance of resonance in concept formation might contemplate the following definition of democracy: *a furry animal with four legs*. This is nonsense, of course. The important point, for present purposes, is that the non-sense of this definition lies in its utter lack of resonance. It violates norms of usage to define "democracy" with the attributes commonly associated with "dog." This is the problem encountered by definitions that are purely stipulative (on the authority of the author). Concepts

[35] Resonance is the criterial embodiment of ordinary-language philosophy. The meaning of a word, declares Wittgenstein (1953: 43), "is its use in the language." Pitkin (1972: 173) expatiates: "The meaning of a word . . . is what one finds in a good dictionary – a word or phrase that can be substituted for it. The meaning of 'justice' has to do with what people intend to convey in saying it, not with the features of the phenomena they say it about." See also Austin (1961); Caton (1963); Chappell (1964); Ryle (1949); Ziff (1960), as well as the various writings of G. E. M. Anscombe, Stanley Cavell, Jerry Fodor, Jerrold Katz, Norman Malcolm, and John Wisdom.

seem arbitrary if they do not fit with established understandings of a term or a phenomenon.

Resonance in the *definition* of a given term is achieved by incorporating standard meanings and avoiding non-standard ones. Resonance in the choice of a *term* is achieved by finding that word within the existing lexicon that (as currently understood) most accurately describes the phenomenon of interest. Where several existing terms capture the phenomenon in question with equal facility – as, for example, the near-synonyms "worldview" and "Weltanschauung" – achieving resonance becomes a matter of finding the term with the greatest common currency. Simple, everyday English terms are more familiar than terms drawn from languages that are dead, foreign, or highly specialized.

Where *no* term within the existing lexicon adequately describes the phenomena in question the writer is evidently forced to invent a new term. Sometimes, neologism is unavoidable, and therefore desirable. Indeed, all words were once neologisms, so we cannot complain too loudly about the forces of innovation. Tradition must occasionally be overturned. That said, one must carefully justify every neologism, every departure from ordinary usage. "The supreme rule of stipulation," writes Richard Robinson, "is surely to stipulate as little as possible. Do not change received definitions when you have nothing to complain of in them."[36]

An example of rather pointless neologism may be drawn from Robert Dahl's work on (as I would say) democracy. Noting the semantic difficulties of this term, and wishing to avoid its "large freight of ambiguity and surplus meaning," Dahl proposed a distinction between democracy, understood as an unattainable ideal, and "polyarchy" (derived from the Greek: *rule of many*), which was to be understood as existing states that exhibit some of the qualities of democracy and are commonly referred to as democracies. This, Dahl thought, would resolve the recurrent tension between "is" and "ought" that embroils the term democracy in scholarly and popular discourse.[37] Dahl's motives are laudable, but one cannot say that the attempted neologism has been successful, despite his prominence in the field. The problem is that the meanings of the two terms are so close that we have trouble hearing polyarchy without thinking of democracy. One might also observe that the attempt to wean social-scientific words from their normative freight is apt to be unavailing, for social science is

[36] Robinson (1954: 80). See also Linnaeus, Aphorisms 243–244 (reproduced in Linsley and Usinger 1959: 40); Connolly ([1974] 1983); Durkheim ([1895] 1964: 37); Mahon (1998); Mill ([1843] 1872: 24); Oppenheim (1975); Pitkin (1972).

[37] Dahl (1971: 9).

generally concerned with things that people have strong feelings about, and these feelings are embedded in ordinary language. Moreover, even if this descriptive–normative division were ultimately successful it would have the unfortunate effect of depriving academic work of popular relevance (Chapter 3). In any case, the key point is that any striking departure from normal usage imposes a cost on the reader of a text. More often than not, this cost is too high and the term is discarded.

Likewise, even the invention of new terms is never entirely removed from the extant lexicon. Neologisms, while rejecting ordinary usage, strive to re-enter the universe of intelligibility. They are rarely nonsense words; they are, instead, new combinations of existing words (e.g., bureaucratic-authoritarianism) or roots (e.g., polyarchy, heresthetic), or terms borrowed from other time periods (e.g., corporatism), other language regions (e.g., equilibrium), or other languages (e.g., laissez faire).[38] By far the most fertile grounds for neologism have been Classical (e.g., Id, communitas, polis, hermeneutics) and eponymous (e.g., Marxism, Reaganism). In all these cases words, or word roots, are imported from their normal contexts to a different context where they take on new meaning or additional senses. However severe the semantic stretch, some original properties remain intact.[39]

To sum up: terms and definitions chosen for use in the social sciences ought to resonate as much as possible with established usage. Inconsistencies with ordinary usage usually introduce ambiguity into a work or a field, despite an author's best intentions. Those concepts that resonate least with ordinary usage may be referred to as neologisms or stipulative definitions; they are excusable only if a more resonant concept is unavailable.

Domain

Granted, all of this depends upon the linguistic terrain within which a concept is expected to resonate. A concept, like an argument, can be evaluated only insofar as its domain of usage is understood. Greater breadth of comprehension and usage is always desirable, all other things being equal. Even so, no social science concept can hope to be truly universal. "Democracy" is understood

[38] On polyarchy, see Dahl (1971); on heresthetic, see Riker (1986); on corporatism, see Collier (1995) and Schmitter (1974).

[39] Robinson (1954: 55) notes: "Men will always be finding themselves with a new thing to express and no word for it, and usually they will meet the problem by applying whichever old word seems nearest, and thus the old word will acquire another meaning or a stretched meaning. Very rarely will they do what A. E. Housman bade them do, invent a new noise to mean the new thing." For a survey of contemporary neologisms, see Algeo (1991).

somewhat differently in different parts of the world.[40] Other terms, such as "vouchers," may have little or no resonance for lay citizens anywhere. Even within the social sciences there are important terminological differences across fields and subfields, and through time. Economists speak a somewhat different language than anthropologists. Consequently, we must be concerned not only with how resonant a concept is, but also with how many language communities it will embrace. There will always be someone, somewhere, who understands a term differently, for whom a proposed definition does not resonate.

Thus, it is important that authors specify – whenever the matter is ambiguous – which language regions a given concept is expected to encompass. Of foremost concern is the distinction between lay and academic audiences. As has been said, it is desirable for social scientists to avoid specialized terms ("jargon") in favor of natural language so that a broader audience can be cultivated for their work. And yet, it must be acknowledged that social science, like all language regions (e.g., medicine, law, street gangs, baseball), requires a specialized vocabulary.[41] Social science cannot accept words simply as they present themselves in ordinary speech. Some fiddling with words and definitions is incumbent on the researcher, if only because ordinary usage is unsettled. Social science concepts, Durkheim points out,

> do not always, or even generally, tally with that of the layman. It is not our aim simply to discover a method for identifying with sufficient accuracy the facts to which the words of ordinary language refer and the ideas they convey. We need, rather, to formulate entirely new concepts, appropriate to the requirements of science and expressed in an appropriate terminology.[42]

The limits of ordinary language as a foundation for social science definition are apparent in the fact that most complex terms – for example, democracy, justice, public goods – carry multiple meanings. Insofar as social scientists need to craft specialized concepts with greater coherence and operationalizability, they are compelled to depart from ordinary usage.

Establishing the domain of a concept depends upon the goals of a piece of research. Sometimes, a general definition – one that travels widely across academic and nonacademic venues – is required. If one is attempting to appeal to policymakers and/or the general public then one must pay close attention to how a given concept will resonate with ordinary usage. If one is attempting to reach beyond a particular culture or language, then usages in other cultures and languages must also be considered. On other occasions, it may not be necessary

[40] Schaffer (1998). [41] Robinson (1954: 73); Sartori (1984). [42] Durkheim ([1895] 1964: 36–37).

to travel widely or to garner universal consensus. This goes for many social science settings, where concepts are crafted for use in a specific project. Here, a more specialized approach to concept formation is warranted – also known as a *stipulative definition*, *definition-in-use*, *contextual definition*, or *systematized concept*.[43]

To illustrate the notion of a conceptual *domain* let us consider the concept of democracy. The domain of this concept may be said to range from a single subfield (e.g., the democratization subfield of political science), to an entire discipline (e.g., political science), to a set of disciplines (e.g., social science), to natural language (e.g., English), or to all natural languages. Each requires a broadening of language communities, and hence (probably) a broader range of definitions and usages that must be encompassed. In order for the concept to function adequately within its domain it must be understood (i.e., resonate) within that domain. This is true regardless of how large, or small, the domain might be.

Just as every concept has a linguistic domain (i.e., the language region where it is intended to resonate) it also has an *empirical* (phenomenal) domain. Consider four contexts in which the concept of democracy is currently employed: (1) local communities; (2) nation-states; (3) trans-national advocacy coalitions; and (4) modes of dress and comportment. Evidently, some attributes are more valid in some of these contexts than in others. For example, "contestation" seems to apply most clearly to (2), and not at all to (4).

In this light, the many definitions of democracy that have been propounded in recent years are not wrong, but rather partial. They explore the meaning of democracy in some contexts while ignoring or downplaying other contexts. They are, in this sense, stipulative, arbitrary – but only if understood as all-purpose definitions. If, instead, we look upon these definitions as limited in domain it becomes possible to restore a modicum of clarity to the vexed enterprise of concept formation.

Consistency

The criterion of *domain* implies the associated criterion of *consistency*. A concept ought to carry the same meaning (more or less) in each empirical context to which it is applied. The range of contexts lying within a concept's population should not elicit different connotations.[44]

[43] Adcock and Collier (2001); Bierwisch (1981); Bierwisch and Schreuder (1992); Robinson (1954); Taylor (1995: ch. 14).

[44] Goertz (2008: 109) calls this "homogeneity."

A violation of consistency – where a term means something different in different contexts – creates a problem of conceptual "stretching."[45] Thus, if corporatism is defined as an institution of peak bargaining among relatively autonomous units within civil society it might be considered a conceptual stretch to extend this concept to include Latin American cases, where unions and other actors in civil society were (and in some cases still are) often manipulated by the state. Of course, if corporatism is defined more broadly – as, say, including any formal bargaining among organized sectors of civil society (with or without state control) – then it does not compromise the concept's integrity to apply it to the Latin American context.

The usual way to adjust the scope of a concept is to add to or subtract from its defining attributes. Usually, one finds an inverse correlation between the intension and extension of a concept. Specifically, when attributes are understood as necessary, necessary-and-sufficient, or additive-and-continuous, adding attributes to a definition diminishes the number of phenomena that satisfy the definition. More focused definitions encompass fewer phenomena. In this manner, an inverse relationship exists between intension and extension, illustrated by the solid line in Figure 5.1.[46]

As an example, let us suppose that we start out with a definition of democracy that includes only the criterion "free and fair elections." Now suppose that we decide to add a second attribute, "civil liberties." If these attributes are understood as necessary or necessary-and-sufficient the addition of each defining trait is likely to narrow the number of polities that qualify as democratic, limiting the extension of the concept. If these qualities are understood as additive and matters of degree (elections are more or less free, civil liberties are more or less respected), the addition of attributes will attenuate the empirical fit between the intension and its extension, in this manner narrowing the empirical boundaries of the concept. (The same set of entities will be viewed as less democratic.) In either situation, the addition of attributes cannot *increase* the extension of a concept, for one is adding definitional requirements.

[45] Collier and Mahon (1993); Sartori (1970).

[46] This relationship is sometimes referred to as a "ladder of abstraction." However, this way of viewing things is somewhat misleading. If democracy is defined by three attributes rather than four it is not more abstract; it simply has a narrower scope (with the caveat noted in the text). In any case, the tradeoff between intension and extension has a long lineage in the literature on logic and concepts. Over a century ago, Stanley Jevons ([1877] 1958: 26) pointed out that when the definitional attributes of a word are expanded – e.g., when "war" becomes "foreign war" – its empirical breadth is narrowed. Weber (quoted in Burger 1976: 72) also noticed that "concepts with ever wider scope [have] ever smaller content." In recent years, this idea has come to be associated with the work of Giovanni Sartori (1970: 1041, 1984; Collier and Gerring 2009). See also Angeles (1981: 141); Cohen and Nagel (1934: 33); Collier and Mahon (1993); Frege (quoted in Passmore [1961] 1967: 184).

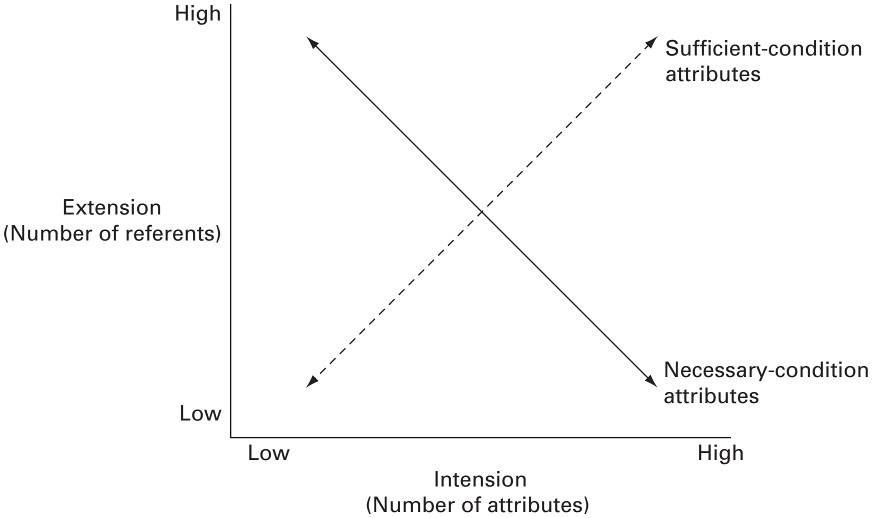


Figure 5.1 Intension and extension: tradeoffs

The utility of this schema is that it allows the conceptualizer to adjust the scope of a concept to fit the needs of an analysis so that violations of consistency are avoided. A concept should be defined so as to "travel" as far as needed, but no further. If one wishes to form a concept of democracy that applies to both Ancient Athens and to the contemporary era, one will need a broader concept than if one is seeking to describe only one or the other. Broadening the concept means choosing a definition that has fewer attributes, and therefore a wider ambit of applicability.

Of course, this tradeoff works differently when defining attributes are understood as sufficient conditions. Here, any addition of attributes *increases* the potential entity space, for each attribute is substitutable for any other attribute.[47] If "contestation" is individually sufficient for a polity to qualify as democratic, then the addition of a second sufficient condition (e.g., "participation") can only increase the population of democracies.[48] Here, we find a direct correlation between intension and extension, illustrated by the dotted line in Figure 5.1.

[47] Goertz (2006).

[48] If the reader feels that this example is forced, one might consider the following. Democracy may be defined generally as rule by the people, with specific dimensions of the concept including: (a) direct popular rule (through referenda and mass assemblies); (b) indirect popular rule (through elected representatives); and (c) deliberative popular rule (through consultative bodies). Arguably, each of the foregoing elements serves as a functional substitute for the others. As such, they may be regarded as sufficient-condition attributes.

It should be recognized, however, that conceptual attributes are rarely understood as sufficient. More typically, they are regarded as necessary-and-sufficient, necessary, or continuous (matters of degree). This means that the tradeoff exemplified by the solid line is more commonly encountered in the work of social science than the tradeoff exemplified by the dotted line. (Further discussion of concept structure is postponed until Chapter 6.)

Fecundity

Social scientists generally associate explanation with causal arguments and understanding with descriptive arguments. However, there is a sense in which descriptive concepts also explain. They do so by reducing the infinite complexity of reality into parsimonious concepts that capture something important – something "real" – about that reality. I shall call this criterion *fecundity*, though it might also be referred to as coherence, depth, fruitfulness, illumination, informative-ness, insight, natural kinds, power, productivity, richness, or thickness. Whatever the terminology, it seems clear that a bid for concepts is a bid to tell us as much as possible about some portion of the empirical world.

Concepts developed by researchers working within the interpretivist tradition often give priority to fecundity. Interpretivists insist that social science cannot evade the call for rich, evocative analysis. Thick description offers advantages over thin description, and thick theories over thin theories: they tell us more about a set of cases. One must appreciate, however, that narrative analysis in and of itself does not ensure fecundity, just as statistical work does not lead inexorably to thin, or reductive, analysis. One can think of many prose artists whose forte is the sweeping generalization, which is neither informative nor evocative. One can think of an equal number of statistical studies that describe or explain a great deal about their subject.[49]

Indeed, qualitative and quantitative methods of concept formation seek the same goal, though by different means. Thus, when systems of biological classification shifted to computer-generated models in the 1960s, resulting classifications were strikingly similar to the existing categories (largely inherited from Linnaeus).[50] Likewise, quantitative explorations of political culture have tended to follow the outline of arguments laid down decades before by Tocqueville, Hartz, and others writing at a time when quantitative analysis was not routinely applied to social questions.[51] Note that the purpose of all descriptive statistical routines

[49] For example, Campbell *et al.* (1960); Verba, Schlozman, and Brady (1995).

[50] Yoon (2009: 202).

[51] Almond and Verba ([1963] 1969).

(e.g., Pearson's *r*, factor analysis, principal component analysis, cluster analysis, and Q-sort analysis) is to elucidate similarities and differences among entities, with the usual aim of sorting them into most-similar and most-different piles. (The same objective applies whether the sorting focuses on cases or on traits.)

Above the level of measurement, the overall goal of a concept might be specified as follows: to focus our attention on some aspect of reality – to pluck it out from the ubiquity of extant data. What makes the concept convincing or unconvincing is the degree to which it "carves nature at the joints" (to use the Platonic metaphor) or identifies "natural kinds" (in Aristotelian language). Concepts strive to identify those things that are alike, grouping them together, and contrasting them to things that are different. Apples with apples, and oranges with oranges.

To be sure, all concepts are on some elemental level conventional. (People are born with the capacity for language, but they are not born with knowledge of a specific language.) However, good concepts move beyond what is merely conventional. They reveal a structure within the realities they attempt to describe. To the extent that a concept manages to identify real similarities and differences it has succeeded in identifying natural kinds. It is ontologically true.

Consider three conceptualizations of regime type. One differentiates between democracies and autocracies;[52] another distinguishes pure democracies, competitive authoritarian states, and pure autocracies;[53] and a third establishes a twenty-one-point index that is intended to function as an interval scale.[54] Which of these is most satisfactory? Evidently, each may be satisfactory for different causal purposes (see below). However, for descriptive purposes the utility of a schema hinges largely upon its fecundity. In the present instance, this means: which schema best describes the subject matter? More specifically, which schema most successfully bundles regime characteristics together, differentiating them from other bundles? Is the natural break-point among regimes to be found between autocracies and democracies (a two-part classification); among pure democracies, competitive autocracies, and pure autocracies; or is there instead a continuum of characteristics with no clear "bundles," justifying a continuous dimensional space? Naturally, many other options might also be considered. Some might argue that regime types are multidimensional, and therefore inappropriate for an ordinal or interval scale.[55] But all such arguments appeal to the ideal of fecundity.[56]

[52] Alvarez *et al.* (1996). [53] Levitsky and Way (2002). [54] Marshall and Jaggers (2007).

[55] Coppedge and Gerring (2011).

[56] A recent quantitative attempt, employing factor analysis, can be found in Coppedge, Alvarez, and Maldonado (2008).

Because of its centrality to concept formation – and to descriptive inference more generally – it is important that we pursue the notion of fecundity in more detail.

Concepts do not make sense unless the attributes that define the concept belong to one another in some logical or functional manner. They must be *coherent*. Within the United States, for example, the concept of "the West" is vulnerable to the charge that western states do not share many features in common (aside from contiguity). Thus, although one can stipulate a precise set of borders (e.g., the seven western-most states) one cannot help but feel that these borders are a trifle artificial. This does not make the concept wrong, but it certainly makes it less meaningful – less fecund – and hence presumably less useful in many contexts. The deeper or richer a concept the more convincing is its claim to define a class of entities deserving of being called by a single name. A coherent term carries more of a punch: it is, descriptively speaking, more powerful, allowing us to infer many things (the common characteristics of the concept) with one thing (the concept's label). The concept of "the South," following the opinion of most historians, would be considered more coherent than "the West," since a much longer list of accompanying attributes could be constructed and differences vis-à-vis other regions are more apparent.

The most coherent definitions are those that identify a core, or "essential," meaning.[57] Robert Dahl, in his influential work on power, sets out to discover "the central intuitively understood meaning of the word," "the primitive notion [of power] that seems to lie behind all [previous] concepts."[58] This essentializing approach to definition is common (and, indeed, often justified). The essential meaning of democracy, for example, is often thought to be rule by the people. This may be viewed as the single principle behind all other definitional characteristics, associated characteristics, and usages of the term. When one says democracy, what one is really talking about is rule by the people. To the extent that this reductionist effort is successful – to the extent, that is, that a single principle is able to subsume various uses and instances of the concept – the highest level of coherence has been achieved in that concept. (Note that essentializing definitions often take the form of minimal definitions, discussed below.)

[57] An "essential," "real," or "ontological" definition is defined as: "Giving the essence of a thing. From among the characteristics possessed by a thing, one is unique and hierarchically superior in that it states (a) the most important characteristic of the thing, and/or (b) that characteristic upon which the others depend for their existence" (Angeles 1981: 57). See also Mill ([1843] 1872: 71); Goertz (2006).

[58] Dahl ([1957] 1969: 79–80).

Differentiation

A concept cannot be internally coherent unless it is distinguishable from other concepts. External differentiation is thus implied by the notion of fecundity. Fecundity refers to how similar a set of phenomena are to each other, while differentiation refers to how different they are from surrounding phenomena. They are flip sides of the same coin. If apples are indistinguishable from oranges, the coherence of "apple" is called into question.[59]

The importance of differentiation is embedded in the words *definition* and *term*. Definition is "the act or product of marking out, or delimiting, the outlines or characteristics of any conception or thing."[60] Term has similar connotations, John Dewey points out. It is "derived from the Latin terminus meaning both boundary and terminal limit."[61] Hanna Pitkin explains, "the meaning of an expression is delimited by what might have been said instead, but wasn't. Green leaves off where yellow and blue begin, so the meaning of 'green' is delimited by the meanings of 'yellow' and 'blue.'"[62] A good concept is, therefore, one with clearly demarcated boundaries.

How, then, does a concept establish clearly demarcated borders? A key element is to specify carefully how a concept fits within a larger semantic field composed of neighboring concepts and referents. We shall refer to this as the background context or *contrast-space* of a concept.

We have noted that concepts are defined in terms of other concepts – boys in terms of girls, nation-states in terms of empires, parties in terms of interest groups. These neighboring terms (synonyms, near-synonyms, antonyms, and superordinate–subordinate concepts) give meaning to a concept. Precisely because of the interconnectedness of language, the redefinition of a term

[59] The twin desiderata of coherence and differentiation correspond to "lumping and splitting" operations in social classification (Zerubavel 1996) and to "similarity and difference" judgments in cognitive linguistics (Tversky and Gati 1978). The twin desiderata may also be recognized in Rosch's work on basic-level categories, which "(a) maximize the number of attributes shared by members of the category; and (b) minimize the number of attributes shared with members of other categories" (Rosch, quoted in Taylor 1995: 50–51).

[60] Reprinted in Chapin (1939: 153). Angeles (1981: 56) traces the Latin origins of the term in the verb "definire," which is translated as "to limit," "to end," "to be concerned with the boundaries of something."

[61] Dewey (1938: 349).

[62] Pitkin (1972: 11). "We call a substance silver," writes Norman Campbell ([1919] 1957: 49), "so long as it is distinguished from other substances and we call all substances silver which are indistinguishable from each other. The test whether a property is a defining or a non-defining property rests simply on the distinction between those properties which serve to distinguish the substance from others and those which it possesses in common with others. Any set of properties which serve to distinguish silver from all other substances will serve to define it."

necessarily involves some resettling of its semantic field. It is impossible to redefine one term without also, at least by implication, redefining others. Any redefinition of corporatism changes our understanding of pluralism, just as a redefinition of democracy changes our understanding of authoritarianism.

It follows – if the meaning of a language is to be sustained – that a new concept should unsettle the semantic field as little as possible, leaving other concepts as they were (more or less).[63] Indeed, a new term or redefinition that poaches attributes from neighboring concepts is laying the ground for future conceptual anarchy. It may resonate on first reading, but is likely to foster confusion in that field or subfield over the longer term. "Crowded" semantic fields are an example of this. Consider the many terms that have been developed over the past several decades to refer to citizen-based groups, including civic association, voluntary association, civil society organization (CSO), citizen sector organization, non-governmental organization (NGO), interest group, and grassroots organization. While subtle differences may be established among these terms it is difficult to accept the endless propagation of terms as productive for the field. Often, neologisms are a sign of conceptual disarray rather than of theoretical fecundity.

In any case, it is incumbent upon writers to clarify how their chosen concept(s) differ from neighboring concepts sharing the same semantic and phenomenal space. This requires establishing clear contrasts with what lies *outside* the boundaries of a concept.

Consider rival concepts seeking to explain American political culture, which may be summarized as *liberalism* (Louis Hartz, Alexis de Tocqueville),[64] *republicanism* (J. G. A. Pocock, Gordon Wood),[65] and a combination of *liberalism*, *republicanism*, and *ascriptive* identities (Rogers Smith).[66] What is of interest here is that these divergent perspectives are often informed by different temporal and/or spatial contrasts. Partisans of the liberal thesis invoke an implicit comparison between the United States and Europe. Partisans of the republican thesis invoke comparisons between the eighteenth and nineteenth centuries – the former being more republican and the latter more liberal. Partisans of the ascriptive thesis invoke comparisons with contemporary ideals and practices – deemed more egalitarian. Each school of thought is probably correct. However, they are correct with respect to different comparisons. American political culture looks different when different temporal and spatial contrasts are invoked.

[63] Sartori (1984). [64] Hartz (1955). [65] Pocock (1975); Wood (1969). See also Shalhope (1972).
[66] Smith (1993).

The same problem of competing contrast-spaces can be observed in many other conceptual debates. For example, writers argue vehemently over the basis of political conflict in contemporary American politics, with some emphasizing the pre-eminence of status, race, and morality[67] and others emphasizing the pre-eminence of social class.[68] (At present, these arguments will be regarded as primarily descriptive rather than causal.) Again, there are many fine points to this debate. That said, it appears that some portion of the disagreement can be explained by contending frames of comparison. Those who hold to the status/values argument may plausibly enlist (a) a spatial comparison with Europe (as did the partisans of the liberal thesis), (b) a temporal comparison with the New Deal era, and (c) a focus on elite-level behavior. Those who hold to the socioeconomic interpretation generally have in mind (a) a temporal comparison that embraces the past half-century (but not Europe or a longer chunk of historical time), (b) mass-level political behavior, and (c) contemporaneous comparisons between the relative strength of status/values issues and class issues in structuring the vote. Again, both schools have plenty of ground to stand on. But it is not the same ground.

Things are similar with respect to recent arguments about global inequality. Those who emphasize the widening gap in global distribution of income tend to base their arguments on evidence drawn from the past several decades, a period when individual-level data is available.[69] Those who emphasize the relative constancy of inequality generally encompass a longer time period – extending back to the mid-twentieth century, and perhaps further.[70] Again, one's conclusions depend critically upon the historical context one chooses to invoke.

Of course, causal arguments also unfold against a contrast-space and this too may create problems, as discussed in Chapter 8.[71] However, it is less likely to engender confusion because the counterfactual is usually more explicit. To say that "*X* causes *Y*" is to say, implicitly, that when *X* changes value, so will *Y* (at least probabilistically). This is fairly well understood, and is formalized in the null hypothesis. But to say that "*Y* is *X*" (i.e., *X*, an adjective, describes *Y*), is to invoke a much more ambiguous contrast-space. "Not *Y*" can refer to *any* temporal or spatial contrast or to the (nonempirical) meaning of the term "*X*" (as in Rogers Smith's argument about American political culture). We are at

[67] Frank (2004); Ladd and Hanley (1975); Morone (2004); Rogin (1987).

[68] Bartels (2006); Fiorina (2005); McCarty, Poole, and Rosenthal (2008).

[69] Milanovic (2005).

[70] Bourguignon and Morrisson (2002); Dollar (2005); Firebaugh (2003).

[71] Achinstein (1983); Garfinkel (1981); Hitchcock (1996); van Fraassen (1980). All work in the "counterfactual" tradition emphasizes this point.

sea, for the null hypothesis – against which the hypothesis might be judged – is not apparent.

Nonetheless, the problem of context becomes tractable insofar as writers are able to address a variety of competing reference points, explicitly and empirically. Of these, there are three possible dimensions: *spatial*, *temporal*, and *conceptual*. The latter, of course, refer to the defining attributes of a concept, and of neighboring concepts. By bringing these comparisons to the fore, virulent arguments, even over highly abstract matters such as political culture and equality, may be joined, and perhaps over time resolved. This is the virtue of explicit comparison, which plays an even more vital role in descriptive inference than in causal inference.

Causal utility

Concepts function causally, as well as descriptively. That is, they serve as components of a larger causal argument. In this latter capacity, they face desiderata that sometimes shape the way they are formed.

For example, suppose one is examining the role of electoral systems in structuring political conflict. Here, one would probably want to limit the ambit of study to polities that are reasonably, or at least minimally, democratic. Consequently, one needs a concept of democracy that achieves this objective. An ideal-type definition (see below) will not suffice; clear borders between democratic and nondemocratic regimes are required. Hence, causal concerns rightly drive concept formation.

In the foregoing example, concepts of democracy demarcate the boundaries of a causal inference. Likewise, concepts also identify causal factors (independent variables) or outcomes (dependent variables). A variable in a causal argument must also function as a concept; there is no such thing as a conceptless variable (if there was, it would lack meaning).

Typically, concepts designed for use as dependent variables group together many attributes. Here, an ideal-type definition may be fruitful. By contrast, concepts designed for use as independent variables are generally smaller, more parsimonious. This fits with the goal of causal argumentation: to explain a lot with a little. It also fits with the goal of causal argumentation to have a clearly defined, discrete "treatment," one that is specific enough to be manipulated (at least in principle) and that can be clearly differentiated from background factors (potential confounders). Additionally, concept formation in the context of causal models must be careful to employ concepts that differentiate a cause from its effect, so that circularity in the argument is avoided.

Of course, concepts defined for use in a specific causal analysis are specialized concepts, not ones that are intended to cover all circumstances and all settings. They are not general in purview. Sometimes, this sort of specialized definition breaks with established usage and thus incurs a cost in the resonance of a concept. This cost must be reckoned with. Causal models are confusing, and impossible to generalize from, if key concepts are defined in idiosyncratic ways.

In sum, causality is only one factor, among many, that rightly affects the formation of concepts (see Table 5.1). Even where the needs of a causal model are pre-eminent, a concept never entirely loses its descriptive purpose. If it did, the causal argument within which it is embedded would lose connection with reality. This is, of course, the very thing of which highly abstract causal models are often accused.[72]

Strategies of conceptualization

Having surveyed general criteria pertaining to concept formation, we turn now to strategies that may help to achieve these goals. Concept formation generally begins with a formal or informal survey of potential concepts. It proceeds by classifying the attributes of each concept so that an overview of each (relevant) concept can be attained. From thence, three general strategies of definition are recommended: *minimal*, *maximal*, and *cumulative*. These sequential strategies are summarized in Table 5.2. The chapter concludes with a brief discussion of the potential utility of this approach for bringing greater order and clarity to the social science lexicon.

Table 5.2 Strategies of conceptualization

1. **Survey of plausible concepts**
2. **Classification of attributes**
3. **Definition**
   - **(a) Minimal** Necessary (and perhaps sufficient) conditions of membership, understood as establishing a minimal threshold of membership.
   - **(b) Maximal** All (nonidiosyncratic) characteristics that define a concept in its purest, most "ideal" form.
   - **(c)Cumulative** A series of binary attributes (0/1) arranged in an ordinal fashion.

[72] Bewley (1999); Hausman (1994); Hedstrom (2005: 3); Maki (2002); Piore (1979); Spiegler and Milberg (2009).

Survey of plausible concepts

Many investigations begin in a frankly inductive mode. There is an empirical terrain of interest – perhaps a community, an institution, or a policy – that becomes the subject of investigation, but without a clear research question or hypothesis. Here, the researcher arrives slowly at a concept, or a set of concepts, to encompass the subject. This is conceptualization in its broadest sense. In this situation, the researcher must canvas widely before settling on a key term(s). Premature closure may cut short the deliberative process by which a subject is processed and understood. Granted, preliminary concepts will always be required; without them, one cannot deliberate at all. However, the canvassing of potential terms – each one treated gingerly, as a hypothesis – is what allows a researcher to test alternative ways of thinking about a topic. What stories are contained in the research site (the archive, the dataset, the ethnographic setting)? Which is the most interesting of these stories? Every story suggests a different label for the project. This is the exploratory process discussed in Chapter 2.

Once the researcher has settled on a preliminary concept he or she ought to briefly review the possible alternatives – that is, the family of near-synonyms that most closely fits the circumstance – resorting to neologism only where absolutely necessary (as discussed above). Since each extant term brings with it a certain amount of semantic luggage, the choice among terms – as well as the choice of how to define the chosen term – rightly involves a canvassing of potential attributes. This step finds precedent in virtually all traditions of conceptual analysis. It is the conceptual equivalent of a "literature review."

Of course, some topics are simple enough to preclude an extensive canvas. Here, recourse to a natural language dictionary or a specialized technical dictionary is sufficient. Alternatively, the author may be able to rely on articles or books that provide a more expanded discussion of a term's meaning and usage patterns, and perhaps its etymology. However, where these short-cuts are unavailing the author will be forced to undertake his or her own conceptual research.

A conscientious semantic canvassing begins with a representative sample of formal definitions and usage patterns for a chosen term, as drawn from relevant scientific fields, from natural language, and from history (etymology). Note that usage patterns may bring to light meanings that are not contained in formal definitions (perhaps because they are so obvious), and may help to clarify meaning when formal definitions are vague. Usage also entails a consideration of the referents of a concept (the phenomena out there to which the concept refers – its extension).

In situations where the different senses of a word are radically disparate – for example, "pen" (writing instrument) and "pen" (enclosure) – one must narrow the conceptual analysis to only one meaning of a term. Of course, homonymy (of which the two radically different meanings of "pen" are an example) and polysemy (where a word invokes a number of closely related meanings) is often a matter of degrees. In borderline cases, the analyst will have to judge which sense should be hived off (to be considered as an independent concept), and which should be retained, so as to create a relatively coherent concept.

Representativeness in the sampling process is achieved by searching for whatever variation in usage and formal definition might exist within a language region and keeping track of the approximate frequency of these various usages and definitions. In future, we may be able to rely on digitized libraries that can be sampled randomly, enabling one to attain a more precise estimate of the frequency of usage and definitional variations. Even so, mechanized sampling will probably not alter our understanding of key terms significantly, for usage patterns within a language region tend to exhibit great regularity. Moreover, our intent is to discard only very idiosyncratic usages and definitions. Thus, as long as the sample is sufficiently broad one is likely to pick up all common (nonidiosyncratic) usages. The principle of redundancy may serve as an indicator of sufficiency: when one reaches a point where definitional attributes and usages begin to repeat, one may justifiably terminate the expedition. One has sampled enough.

The issue of linguistic domain – how many language regions to survey – is also crucial. A sampling is better if it covers more language regions. Yet if this broad search reveals significant differences in meaning then the analyst may restrict the scope of the investigation in order to preserve consistency and coherence. Any sampling is likely to have a home turf – perhaps a particular field of social science – that is extensively canvassed, and other areas that are surveyed more superficially. In any case, the domain of the survey will help to establish the domain of the resulting definition.

Classification of attributes

The next task is to reduce the plenitude of meanings implied by a term into a single table. The construction of such a table rests on the assumption that, although definitions for a given term are, in principle, infinite (since even a small number of attributes can be combined in many ways, and since there are always multiple ways to convey a similar meaning), most definitions and usages juggle the same basic set of attributes. By combining near-synonyms and by organizing them along different dimensions one ought to be able to reduce the definitional profusion of even the most complex concept into a relatively parsimonious table of attributes. We regard this table as the lexical definition of a term because it reports the many meanings of that term extant across a given linguistic domain.

As an example, let us explore the definitional attributes of "democracy." Our survey of definitions and usages rests on a number of recent studies that attempt to delineate the meaning of this key term, focusing primarily on the Western tradition (historical and contemporary).[73] This is therefore regarded as the principal domain of the concept. Empirically, I choose to focus on applications of this concept within political contexts, and especially in large polities such as the nation-state (rather than within small, local bodies). This will be the empirical domain of the concept. From this compendium of definitions and usages, one may distill a list of common attributes, depicted in Table 5.3. Obviously, this list rests at a fairly abstract level; one could extend it to include much more specific features of the political landscape. But this would require a much larger table and is unnecessary for present purposes.

With a complex subject like democracy it is helpful if the attributes can be arranged in a taxonomic fashion (Chapter 6). Of course, this is not always possible, and one can glimpse more than a few violations of taxonomic principles (e.g., components that traverse several categories). Still, this exercise in semantic reduction is useful wherever practicable.

Definition: concept types

With the caveats noted above, it seems fair to regard Table 5.3 as a fairly encompassing lexical definition, including most of the attributes commonly associated with the term in the Western tradition. Even so, because of the number and diversity of these attributes, Table 5.3 does not take us very far toward a final definition. In order to create a more tractable empirical concept, one must go further. This next step – from lexical definition to specialized definition – is crucial. To achieve it, three approaches will be reviewed: *minimal*, *maximal*, and *cumulative*.

[73] Beetham (1994, 1999); Collier and Levitsky (1997); Held (2006); Lively (1975); Sartori (1962); Saward (2003); Weale (2007).

Table 5.3 A classification of fundamental attributes: "Democracy"

Core principle: rule by the people	
I Electoral (aka elite, minimal, realist, Schumpeterian) *Principles*: contestation, competition. *Question*: are government offices filled by free and fair multiparty elections? *Institutions*: elections, political parties, competitiveness, and turnover.	**II Liberal** (aka consensus, pluralist) *Principles*: limited government, multiple veto points, horizontal accountability, individual rights, civil liberties, transparency. *Question*: is political power decentralized and constrained? *Institutions*: multiple, independent, and decentralized, with special focus on the role of the media, interest groups, the judiciary, and a written constitution with explicit guarantees.
III Majoritarian (aka responsible party government) *Principles*: majority rule, centralization, vertical accountability. *Question*: does the majority (or plurality) rule? *Institutions*: consolidated and centralized, with special focus on the role of political parties.	**IV Participatory** *Principle*: government by the people. *Question*: do ordinary citizens participate in politics? *Institutions*: election law, civil society, local government, direct democracy.
V Deliberative *Principle:* government by reason. *Question*: are political decisions the product of public deliberation? *Institutions*: media, hearings, panels, other deliberative bodies.	**VI Egalitarian** *Principle*: political equality. *Question*: are all citizens equally empowered? *Institutions*: designed to ensure equal participation, representation, protection, and politically relevant resources.

Institutions: both governmental and nongovernmental (e.g., interest groups, parties, civic associations).
Source: Coppedge and Gerring (2011).

Minimal

One long-standing definitional strategy seeks to identify the bare essentials of a concept, sufficient to differentiate it extensionally without excluding any of the phenomena generally understood as part of the extension. The resulting definition should be capable of substituting for all (nonidiosyncratic) uses of the term without too much loss of meaning. This means, of course, that it should not conflict with any (nonidiosyncratic) usages. Each attribute that defines a concept minimally is regarded as a necessary condition: all entities must possess this attribute in order to be considered a member of the set. Collectively, these attributes are jointly sufficient to bound the concept extensionally. Minimal definitions thus aim for crisp borders, allowing for the classification of entities as "in" or "out." Of course, they may not always achieve this goal, but this is their aim.[74]

Sometimes, minimal concepts are crafted around an abstract core principle such as "rule by the people." In this instance, the core meaning satisfies the criterion of resonance, for all invocations of democracy revolve in some way around this idea. However, such an abstract definition does not achieve crisp borders for the concept; indeed, it scarcely identifies borders. In this respect, it is problematic.

A more common approach is to identify a specific component of the term that everyone (or nearly everyone) agrees upon. If we are limiting ourselves to representative polities (excluding direct democracies) one might argue that free and fair elections constitutes a necessary condition of democracy. This attribute suffices as a minimal definition, for it is sufficient to bound the entity empirically. That is, having free and fair elections makes a polity a democracy; no other attributes are necessary. At least, so it might be argued.

The caveat, of course, is that we are defining democracy in a very minimal fashion, leaving other attributes often associated with the concept in abeyance. This imposes some costs in resonance. The stripped down meaning of the term sounds strange to those attuned to democracy's many nuances.

Maximal

Maximal definitions, in contrast to minimal definitions, aim for the inclusion of all (nonidiosyncratic) attributes, thereby defining a concept in its purest, most "ideal" form. This would, of course, include the attribute(s) that defines the concept minimally: its necessary condition(s). As Weber describes it, "an ideal-type is formed . . . by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent *concrete individual* phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified *analytical* construct."[75]

Following this recipe, one might create an ideal-type definition of democracy that includes most, or all, of the dimensions listed in Table 5.3. Of course,

[74] Definitional strategies similar to the "minimal" strategy have been employed by various writers, although not usually by this name. See, e.g., Debnam (1984) on "power"; Freeden (1994: 146) on "ineliminable" attributes; Hamilton (1987) on "ideology"; Pitkin (1967: 10–11) on "basic meaning"; Murphey (1994: 23–24). Sartori endorses minimal definition in early work (1975: 34–35, 1976: 61), but drops the matter in his classic work on concept formation (1984). It should be noted that minimal definition is similar, though not identical, to a "procedural minimum" definition (Collier and Levitsky, 1997). In the latter, the search is for an operationalization that satisfies all definitional requirements of a concept.

[75] Weber ([1905] 1949: 90). See also Burger (1976). In citing Weber, I do not claim to be using the concept of an ideal-type in precisely the way that Weber envisioned.

some might be excluded if it could be argued that they detract significantly from the coherence of the overall concept. Blatantly contradictory elements should be avoided.

Ideal-types, as the term suggests, need not have a specific real-life empirical referent. Perhaps no extant polity achieves perfect democracy. However, in order to be of service an ideal-type must approximate real, existing entities, which are then scored according to how closely they resemble the attributes of the ideal-type. Ideal-types are always matters of degree, and hence generally operationalized by interval scales (discussed in Chapter 6).

Cumulative

A third strategy of concept formation is an attempt to reconcile minimal and maximal approaches by ranking the (binary) attributes commonly associated with a concept in a cumulative fashion, that is, as more or less essential to a concept.[76] This results in an ordinal scale (discussed in Chapter 6).

Following these principles, one can envision a cumulative scale indicator of democracy that begins with free and fair elections – the minimal definition – and proceeds through eight additional criteria, listed in order of centrality to the concept of interest, as depicted in Table 5.4. If this ordering of attributes is accepted – if, that is, it is agreed that 1 is more essential than 2 and 2 is more essential than 3 – then it may be possible to arrive at an acceptable definition of democracy that incorporates many of the attributes commonly associated with the term, while also recognizing the relative importance of each of these attributes. It has the additional advantage of allowing us to order all extant polities empirically according to their degree of democracy: the more attributes a polity possesses, the more democratic it is.[77] (This solves the aggregation problem, an issue of measurement discussed in Chapter 6.)

Of course, we will not be able to determine *how much* more democratic one polity is than another, for we cannot presume that each level is equidistant from the next (the distinction between an ordinal and interval scale). A second shortcoming of this particular cumulative definition is that the ordinal scale of attributes may not be fully comprehensive; some attributes may be difficult to rank in terms of their centrality to the concept. Indeed, one can see that not all of democracy's lexical attributes (see Table 5.3) are contained in the cumulative concept in Table 5.4.

[76] This is very similar in spirit to the construction of a Guttman scale, except that we are dealing with attributes rather than indicators, and with the theoretical (rather than empirical) properties of these attributes.

[77] For another example of the ordinal technique see Coppedge and Reinicke (1990).

Table 5.4 Cumulative definition: "Democracy"

	Ordinal scale								
Attributes	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
(a) Free and fair elections	x	x	x	x	x	x	x	x	x
(b) Self-government (domestic)		x	x	x	x	x	x	x	x
(c) Self-government (complete)			x	x	x	x	x	x	x
(d) Executive elected and paramount				x	x	x	x	x	x
(e) Universal male suffrage					x	x	x	x	x
(f) Universal suffrage						x	x	x	x
(g) Executive constitutionality							x	x	x
(h) Executive constraints								x	x
(i) Civil liberty									x

(a) Free and fair elections: national elections are regularly held, are open to all major parties and candidates (including all opposition parties and figures who might pose a significant challenge to the ruling group), and appear on balance to reflect the will of the electorate (whatever irregularities might exist).
(b) Self-government (domestic): sovereignty over domestic policy.
(c) Self-government (complete): sovereignty over domestic and foreign policy.
(d) Executive elected and paramount: executive is elected and is paramount (i.e., superior, *de facto*, to other leaders and institutions).
(e) Universal male suffrage: all adult male citizens are allowed to vote and no group of citizens is selectively discouraged from voting. Presumption: citizenship includes a majority of permanent residents in a territory.
(f) Universal suffrage: all adult citizens are allowed to vote and no group of citizens is selectively discouraged from voting. Presumption: citizenship includes a majority of permanent residents in a territory.
(g) Executive constitutionality: executive acts in a constitutional manner, and does not change the constitution to suit its political needs (though it may try).
(h) Executive constraints: executive, although paramount, is effectively constrained by other political institutions, acting in their constitutional role (e.g., judiciary, legislature, monarch, independent agencies).
(i) Civil liberty: citizens enjoy freedom of speech and freedom from politically motivated persecution by government.

Discussion

Having outlined three strategies of concept definition – minimal, maximal, and cumulative – the reader may wonder whether this exhausts the field. Naturally, it does not. Concepts serve many theoretical and empirical functions, and these functions rightly condition how they are formed within the purview of a given work. However, *general* definitions of a concept – those intended to travel widely – tend to adopt minimal or maximal approaches to definition. (Occasionally, they may employ a cumulative approach.) This is because these approaches tend to be most successful in establishing resonance, consistency, and coherence across a broad domain. (Issues of measurement are generally secondary when a concept must travel widely.) In other words, minimal and maximal definitions offer a better resolution of the criterial demands that all concepts face (see Table 5.1).

To be sure, some concepts resist this effort at semantic reduction. It is alleged that some concepts embody "family-resemblance" attributes, where different usages share no single characteristic in common and therefore have no core meaning. An oft-discussed example is "mother," which may be defined as (a) a biological fact, (b) the person who plays a principal role in nurturing a child, or (c) according to rules and norms within specialized domains (e.g., Mother Superior within the Catholic hierarchy). These definitions share no single element in common. They are disparate.[78]

In social science context, however, we are less likely to witness family-resemblance concepts. Democracy is an essentially contested concept. Even so, all commentators seem to agree that, as applied to political contexts, this concept revolves around a single core attribute – rule by the people. "Justice," another bone of contention, also has a core meaning: to each his or her due. (As it happens, both of these core meanings can be traced back to Ancient Greece.)

More to the point, even in situations where family resemblances might be said to exist there is little profit in trumpeting the disparate nature of a term's definitions. Thus, while "corporatism" has been regarded as a family-resemblance concept[79] it could also be subjected to a minimal or maximal definition. I would argue that we are better served by the latter than by the former precisely because minimal and maximal definitions create more coherent concepts, and ones that are easier to locate in empirical space (i.e., to measure), albeit with some loss of resonance. Better a minimal, maximal, or cumulative definition that is flawed – as in some sense, all social science definitions are – than a family-resemblance definition that results in an incoherent concept.

Before concluding it is worth taking note of the fact that we have focused thus far on "hard" cases – democracy, justice, and the like. Other concepts in the social science lexicon are rarely as troublesome. From this perspective, the problem of conceptualization is perhaps somewhat less severe than it may seem from a cursory reading of this chapter.

By way of contrast, let us quickly examine an easier, more concrete concept. "Political party" may be defined minimally as an organization that nominates individuals for office. This definition imposes crisp borders and is substitutable for all extant usages of which I am aware. A maximal definition would,

[78] Wittgenstein (1953). See also Collier and Mahon (1993); Goertz (2006); Taylor (1995: ch. 3).

[79] Collier and Mahon (1993: 847).

of course, encompass other attributes commonly associated with the work of political parties, such as a shared ideology, an organizational apparatus, well-defined membership, and endurance over time. These attributes describe parties in their strongest, most ideal sense, and are matters of degree. A cumulative definition would arrange these same attributes (or some subset of them) according to their centrality to the concept.[80] Whichever strategy one chooses to employ, defining "political party" is considerably easier than defining "democracy." And so it may be for other concepts that lie closer to the empirical bone.

Even with the most complex concepts, carefully crafted definitions in the minimal, maximal, or cumulative mold should provide a common scaffolding upon which the work of social science can rest in a reasonably stable and consistent manner. To be sure, meanings change over time; but such change occurs slowly. New terms, or new meanings for old terms, appear idiosyncratic at first. Over time, if neologisms gain adherents, they become established. However, *at any given point in time* reasonably authoritative definitions should be feasible – with the caveat that multiple approaches to the same concept (minimal, maximal, and cumulative) can often be justified.[81] Thus, it is incumbent upon authors to clarify what style of definition they are adopting.

Note also that the construction of minimal and maximal definitions establishes semantic *boundaries* around a concept. It specifies the minimal and maximal attributes, and the corresponding minimal and maximal extensions. This sort of exercise – equivalent to an "extreme bounds" analysis – is especially useful when dealing with far-flung concepts such as democracy.

[80] For further discussion of this concept see Gunther and Diamond (2003: 172).

[81] For further discussion and additional examples, see Gerring (1997); Gerring and Barresi (2003).