Nate Pavitt
Nashville
2016

Experimental Political Science and the Study of Causality

From Nature to the Lab

REBECCA B. MORTON
New York University

KENNETH C. WILLIAMS
Michigan State University

CAMBRIDGE
UNIVERSITY PRESS

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

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521136488

First published 2010
Reprinted 2012

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Morton, Rebecca B., 1954–
Experimental political science and the study of causality : from
nature to the lab / Rebecca B. Morton, Kenneth C.Williams.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-19966-7 – ISBN 978-0-521-13648-8 (pbk.)
1. Political science – Methodology. 2. Political science – Research.
3. Thought experiments. I. Williams, Kenneth C. II. Title.
JA71.M675 2010
320.01–dc22 2010019826

ISBN 978-0-521-19966-7 Hardback
ISBN 978-0-521-13648-8 Paperback

Contents

2

Experiments and Causal Relations

2.1 Placing Experimental Research in Context

In typical discussions of estimating causality, social scientists who c
from a statistics perspective often begin with a review of the experime
approach in an idealized setting that rarely exists, argue that the exp
mental approach as idealized is not feasible in social science, and the
on to discuss how causality is measured in observational data. For exan
Winship and Morgan (1999) begin their otherwise excellent review o
literature in social science on measuring the effects of causes with the s
ment (p. 660), "sociologists, economists, and political scientists must
on what is now known as observational data – data that have been ge
ated by something other than a randomized experiment – typically surv
censuses, or administrative records." This tendency to bracket off mea
ing causality in experimental social science from measuring causalit
observational data presumes that experiments are "either-or propositio
a researcher can either conduct an "ideal" experiment, which we argu
this book would not be ideal for many questions in which political scien
are interested, or work with observational data.

Most of experimental social science is not the hypothesized idea
classical experiment, usually with good reason. The bracketing off prev
a discussion of how causality is measured in experiments as they exis
social science and a realistic comparison of those methods to research w
observational data. Moreover, many of the methods that are used to mea
causality in observational data are also relevant for experimental wor
social science, and researchers must understand the relationships betw
experimental design and these methods and how they interact.

As discussed in Section 1.5.1, if you ask political scientists what is
principal advantage of the experimental approach to political science, n

would answer that it can better measure causality than is possible with observational data. Yet, the relationship between the experimental approach to establishing causality and other methodological approaches to causality used in political science is not well understood and misunderstandings exist between experimentalists whose approach builds on work in social psychology or statistics and those whose approach builds on work in economics over how causality can be measured in experiments. A significant source of this lack of a common understanding is related to the welcoming-discipline nature of political science, as mentioned in Section 1.6.

Both the advantages and the disadvantages of being a welcoming discipline are also exhibited in the ways in which political scientists have addressed questions of causality. Political scientists rely on approaches to causality, which originated in statistics and biostatistics, sociology, psychology, and economics. This borrowing has benefited political scientists' research as the discussion of particular examples in Chapter 1 demonstrates. However, little has been written about how these approaches fit together or make sense for political science questions in a comprehensive sense (Zeng [2003] is exception) or how these approaches compare to the experimental approach. In this chapter, we explore how causality is measured (or not) in political science research and place the experimental approach within that context.

The importance of a discussion of measuring causality in a general sense for both experimental and observational data was highlighted by the recent exchange between Imai (2005) and Gerber and Green (2000) in the *American Political Science Review* on how to interpret the estimated effects in field experiments on mobilization. We thus discuss measuring causality in a general sense for a given data set, not making any assumptions about whether the data are observational or experimental. From this general perspective we then place the experimental approach to measuring causality – its advantages and disadvantages compared to observational data – in context.

2.2 Technical Background

The statistical and econometric literature on causality is built on existing knowledge in probability theory, procedures such as ordinary least squares, methods such as probit and logit, the maximum likelihood approach to data, nonparametric estimation techniques, and graph theory. As such, some of the material in this and following chapters necessarily refers to these techniques and for the more technical sections we assume that readers have prior exposure to probability theory and basic multiple regression techniques and

can work with simple mathematical models.[1] It is necessary to present t technical material because of the importance of the assumptions under ing the techniques for how we interpret data on causality, even assumptio that to an outside observer would appear to be innocuous. We attempt make the material accessible to readers who have less exposure to the techniques in our interpretations.

Two important caveats about our discussion of causality are in order. Fir we only discuss quantitative approaches to causality because our emph sis is on the measurement of causal relations through the analysis of da generated by manipulations either by an experimentalist or nature acti in a similar fashion. We do not address qualitative approaches to causal used in political science. Second, we primarily focus on studies of causal using cross-sectional and panel data rather than time-series data becau most empirical studies with experimental data involve the use of such da Specifically, when we have observations over a period of time, we are inte ested in cases for which it is reasonable to argue that it is more likely th the number of observations per time period approaches infinity faster th the number of time periods and panel data methods are appropriate.[2]

2.3 Causes of Effects Versus Effects of Causes

2.3.1 Causes of Effects and Theoretical Models

When addressing causality we must distinguish between investigations the causes of effects and investigations of the effects of causes. If we a asking what causes turnout, we are asking a question about the causes effects (turnout), but if we ask if making a voter more informed increas his or her probability of voting, then we are asking more narrowly abo the effects on turnout of a cause (information). Heckman (2005a) preser a view from econometrics when he argues that ultimately we are inte ested in the causes of effects. He remarks (p. 2): "Science is all about co structing models of the causes of effects." Heckman also (2005b) conten that

> "causality" is not a central issue in fields with well formulated models where it usual emerges as an automatic by-product and not as the main feature of a scienti investigation. Moreover, intuitive notions about causality have been dropped

[1] In some discussions, knowledge of logit and probit is helpful, although not required.
[2] Hence, the asymptotic properties of estimators are evaluated as the number of observatio approach infinity, holding time constant. In time-series data, the opposite is the case.

pursuit of a rigorous physical theory. As I note in my essay with Abbring (2007), Richard Feynman in his work on quantum electrodynamics allowed the future to cause the past in pursuit of a scientifically rigorous model even though it violated "common sense" causal principles. The less clearly developed is a field of inquiry, the more likely is it to rely on vague notions like causality rather than explicitly formulated models.

The emphasis on models of causes of effects as the primary goal of study is no doubt the main reason why Heckman advocates what he calls the structural approach to causality, which with observational data is close to the formal theory approach and which we explore in detail in Chapter 6.

In the formal theory approach to causality, an empirical researcher works with a model of the causes of effects from previous theoretical and empirical work and then evaluates that model (predictions and assumptions) with available data, either observational or experimental. The model usually makes a number of causal predictions rather than just one, but all are logically consistent with each other and with the model's assumptions. The causality in the model is often conditional to given situations; that is, some variables may be simultaneously determined. The evaluation of the model leads to further research, both theoretical and empirical. Sometimes theoretical investigators may think like Feynman; that is, envision situations that are beyond common sense in order to explore the logical implications of the model in these nonsensical worlds. Empirical investigations, however, tend to use applied versions of the model (although experiments can allow for the researcher to move beyond the observed world in the same way theory allows, if the researcher desires). This approach is also presented in political science by Morton (1999) and Cameron and Morton (2002) and is the basis of most laboratory experiments conducted by political economists and some by political psychologists (although with nonformal rather than formal models).

The weight on modeling the causes of effects in economics explains why many experimentalists who come from an economics tradition do not appear to be terribly interested in using their experiments to study a particular single cause-and-effect relationship in isolation but instead typically study a host of predicted relationships from some existing theory, as discussed in Chapter 6. These experimentalists usually begin with a formal model of some process, derive a number of predictions from that model, and then consider whether the behavior of subjects is in line with these predictions (or not) in their experiment. To researchers who have been trained to think of experiments as single tests of isolated cause-and-effect relationships as in the so-called classical experiment, these experiments

appear wrongheaded. But this failure is one of understanding, not o
method, which we hope our discussion of the formal theory approach
causality in this book will help reduce.

2.3.2 Effects of Causes and Inductive Research

However, not everyone agrees with Heckman's emphasis on theoret
models and the causes of effects. In his critique of Heckman's essay, So
(2005, p. 103) argues that many scientific questions are not causal, but pu
descriptive. He remarks that "NASA . . . crashed a probe from the D
Impact spacecraft into comet Tempel1 with the objective of learning m
about the structure and composition of cometary nuclei." Sobel contin
by pointing out that modeling the causes of effects is not important un
the effects of causes are sizable, noting that studying the causes of glo
warming is important because of the effects of global warming.

A lot of political science quantitative research – we would say the mo
approach – is not so much into modeling or thinking beyond causality
instead focuses on investigating the effects of particular causes. Someti
this activity is advocated as part of an effort to build toward a general mo
of the causes of effects, but usually if such a goal is in a researcher's min
is implicit. In experimental research, Gerber and Green (2002) advocate
approach in their call for use of field experiments to search for facts, as
discuss further later. Gerber and Green contend that experiments are a p
ticularly useful way to discover such causal relationships, more useful t
research with observational data. Experimentalists who have been lar
trained from a statistical background and some political psychologists
tend to take this approach. The implicit idea is that eventually system
reviews would address how these facts, that is, causes, fit together and l
us understand the causes of effects.

Is there a "right" way to build a general model of the causes of effe
Morton (1999) maintains, as do we, that both approaches help us b
general models of the causes of effects. Moreover, as Sobel holds, someti
purely descriptive studies, which are not interested in causal questions,
useful. But it is a mistake to think that piecemeal studies of the effect
causes can be effectively accomplished without theorizing, just as it
mistake to think that general models of the causes of effects can be b
without piecemeal studies of effects of causes in the context of the mo
To make this point, we explore how piecemeal studies of the effects of ca
and approaches to building models of the causes of effects work in this
the following chapters.

2.3.3 An Example: Information and Voting

To illustrate how causal inference in political science research is conducted, we focus on a research area that has received significant attention, using both observational and experimental data and from researchers who use methods from a range of disciplines: What is the causal effect of information on voting behavior? What are the causes that determine how individuals vote in elections? We later elaborate on the nature of the research questions in terms of causality.

The Effects of a Cause Question

Elections often involve fairly complicated choices for voters. Even in simple two-candidate contests, voters vary in the degree over which they know the policy preferences of the candidates and how the candidates are likely to govern if elected. When elections involve more than two candidates or are referenda over specific legislation, voters' information about the consequences of their choices also varies. What is the effect of information about choices in elections on how voters choose? We know that uninformed voters are more likely to abstain; Connelly and Field (1944), in one of the first survey analyses of the determinants of turnout, found that nonvoters were two-thirds more likely to be uninformed about general political matters than those who participated. But, as Connelly and Field noted, the effect they discovered may simply reflect the fact that nonvoters are also less educated. Connelly and Field could not conclude that a lack of information caused voters to abstain. Much subsequent research has reported that this relationship is robust across election contests and years of study. Are these individuals not voting because they are less educated and, as a consequence, choosing to be uninformed because they are not voting or are they uninformed because they are less educated and, as a consequence, choosing not to vote? Or is there another factor, such as cognitive abilities or candidate strategies, that affects both whether someone is informed and whether they vote or not?

Furthermore, what is the effect of information on voting choices if voters do participate? Some uninformed individuals do vote. Do less informed voters choose differently than more informed voters who are similar in other ways, choosing different candidates as Bartels (1996) contends? Or do uninformed voters choose "as if" they are informed using simple cues like party labels or poll results, as argued by a number of scholars?[3] How much information do voters need to make "correct" decisions (decisions they would make if they were fully informed)? Can voters use simple cues

[3] See, for example, Berelson et al. (1954); McKelvey and Ordeshook (1985); Page and Shapiro (1992).

and cognitive heuristics as described by Kahneman et al. (1982) to ma
"correct" decisions? If uninformed voters would choose differently if th
were fully informed, then does the distribution of information affect t
ability of different voters to have their preferences affect electoral outcom
resulting in election outcomes that are biased in favor of the preferenc
of other, more informed voters? The answers to these questions are func
mental for understanding how electoral processes work and how electio
translate voter preferences into outcomes. All of these answers hinge
how information influences voter choices, a question that turns out
be extremely difficult to determine and the subject of much continui
controversy.[4]

The Causes of an Effect: Questions and Theories of Voting

Furthermore, the relationship between information and voting is high
relevant to the task of building a general model of turnout and voti
behavior in elections. Why do people vote? What determines how they vo
There are a number of competing explanations for voting behavior, most
which have specific implications for the relationship between informati
and voting. We explore the main ones because they are relevant for some
the examples that we use throughout this text.

The Expressive Voter. One explanation of how voters choose in elections
that voters choose whether to participate and how they vote for expressi
purposes, which we label the Expressive Voter Theory.[5] Voters receive son
value from participation and expressing their sincere preferences, and th
induces them to do both. A version of this theory argues that one implicatio
is that the more informed voters are, the more they are likely to participate
elections because they receive more utility from expressing preferences whe
they are informed about the choices.[6] Expressive voters are also predicte
to make the choice that their information leads them to believe is ex an
their most preferred choice.

The Cognitive Miser. An explanation of how voters choose from politic
psychology is the view that voters are "limited information processors"
"cognitive misers" and make voting decisions based on heuristics and cu

[4] Contrast, for example, the conclusions of Bartels (1996), Lau and Redlawsk (2001), ar
Sekhon (2005) on whether uninformed voters vote "as if" they are informed and th
literature reviewed on this subject. We address the reasons for these different conclusio
subsequently.

[5] See Schuessler (2000), for example.

[6] See Matsusaka (1995).

as described earlier. These heuristics may lead to more informed choices with limited information or they may lead to systematic biases in how voters choose. As Lau and Redlawsk (2001, p. 952) remark: "Heuristics may even improve the decision-making capabilities of some voters in some situations but hinder the capabilities of others." Thus, the theory contends that how voters use these cognitive heuristics and whether they can lead to biased outcomes influences how information affects voters' choices. We label this the Cognitive Miser Theory.

The Primed, Framed, or Persuaded Voter. An extension of the Cognitive Miser Theory is the view that in politics, because voters are cognitive misers they can be easily influenced by information sources such as campaign advertising and the news media. That is, as Krosnick and Kinder argue, one heuristic that voters might use "is to rely upon information that is most *accessible* in memory, information that comes to mind spontaneously and effortlessly when a judgement must be made" (1990, p. 499, emphasis in the original). Because information comes to voters selectively, largely through the news media or advertising, biases in this information can have an effect on voter behavior. The contention is that the news media, by choosing which stories to cover and how to present the information, can "frame" the information voters receive, "prime" them to think about particular issues, or "persuade" voters to value particular positions, such that they are inclined to support political positions and candidates.

Chong and Druckman (2007) review the literature on framing and explain the distinctions among framing, priming, and persuasion as used in the psychology and communications literatures. Loosely, framing effects work when a communication causes an individual to alter the weight he or she places on a consideration in evaluating an issue or an event (e.g., more weight on free speech instead of public safety when evaluating a hate group rally), whereas priming in the communication literature refers to altering the weight attached to an issue in evaluations of politicians (e.g., more weight on economic issues than on foreign affairs in evaluating the president). Persuasion, in contrast, means changing an actual evaluation on a given dimension (e.g., the president has good economic policies). Thus, the theory argues that biases in the content of the information presented to voters and differences in presentations of the information can bias how voters choose in elections.

Effects of Negative Information. One particular aspect of information during election campaigns has been the subject of much disagreement in the

political behavior literature – the effects of negative campaign advertis
Ansolabehere and Iyengar (1997) suggest that some advertising can a
ally decrease participation. Specifically, they argue that negative adverti
actually demobilizes voters by making them apathetic. The exposure to n
ative advertising, according to this theory, weakens voters' confidence in
responsiveness of electoral institutions and public officials generally.
negative advertising suggests not only that the candidate who is the sub
of the negative ads is not someone to trust, but also that the political sys
in general is less trustworthy. Negative advertising then makes voters m
negative about politics, more cynical, and less likely to participate. In c
trast, others such as Lau (1982, 1985) have argued that negative adverti
actually increases voter participation because the information provided
be more informative than positive advertising. The debate over the eff
of negative advertising has been the subject of a large experimental li
ature in political science and is also a case for which a notable num
of observational studies exist that use experimental reasoning. We disc
some examples from this literature.

The Pivotal Voter. An alternative theory of voting from political econon
is what we label the Pivotal Voter Theory. In this model, voters' choi
whether to turn out and how to vote, are conditioned on being pivotal. T
is, whether or how an individual votes does not matter unless his or her
is pivotal. So when choosing whether and how to vote, an individual v
"as if" he or she is pivotal and does not vote at all if the expected benefits fi
voting (again conditioned on pivotality) are less than the cost. In a sem
set of papers, Feddersen and Pesendorfer (1996) apply the pivotal v
model to understand how information affects voters' choices. They sh
that the theory predicts that uninformed voters may be less likely to vote t
informed voters if they believe that informed voters have similar preferen
because they wish to avoid affecting the election outcome in the wr
direction. Moreover, the less informed voters may vote to offset parti
voters whose votes are independent of information levels. According to
theory, then, it is possible that less informed voters may purposely
against their ex ante most preferred choices to offset the partisan vot
These particular predictions about how less informed voters choose
been called by Feddersen and Pesendorfer the Swing Voter's Curse.

The Voter as a Client. Electoral politics in many developing countries
been theorized by comparative politics scholars as a clientelist system. Cli
telism is when the relationship between government officials and voter

characterized as between a rich patron who provides poor clients with jobs, protection, and other specific benefits in return for votes. Thus, in such systems, campaign messages are about the redistributive transfers that the elected officials plan to provide to their supporters. Voters choose candidates in elections that they believe are most likely to provide them with the most transfers. Information about what candidates will do once in office in terms of such transfers can thus affect voters' choices to the extent that they value the transfers.

Of course, because voting is a fundamental part of political behavior and has been the subject of extensive theoretical examination, other theories exist of how people vote, such as group models of voting described by Feddersen and Sandroni (2006), Morton (1987, 1991), Schram (1989), and Uhlaner (1989). We focus on the aforementioned theories because they have been addressed using experimental work that we use as examples in this chapter.[7]

The Broader Implications

Evaluating the causal effect of information on turnout and how individuals vote in the ballot booth provides evidence on whether these particular implications of the more general models of the causes of voting are supported. Such research, combined with evaluations of other implications of these theories, works to determine what causes turnout and what causes how voters choose in the ballot booth.

Furthermore, the answers to the questions of effects of a cause and the causes of an effect also affect how we answer other important policy questions about elections and campaigns. For example, how do campaign advertisements influence voters' choices (if at all)? Do ads need to be substantively informative to influence uninformed voters to choose as if they are informed or can voters use simple ads that mention things like party or other simple messages to make "correct choices?" Is it important that the media provide detailed substantive information on candidate positions? Can biased media reporting on candidate policy positions influence voters? How important are debates in which candidates discuss substantive issues in the electoral process? These policy questions depend not only on the particular causal effect of information on voting but also how we answer the questions about why voters turn out and the determinants of how they vote.

These questions are also useful for an exploration of how causality is investigated in political science using both experiments and nonexperimental

[7] Feddersen et al. (2009) provide an interesting experimental test of a theory of voting related to the Feddersen and Sandroni model of ethical voting.

empirical studies since many researchers have tackled them using bo types of data, including even natural experiments. Thus, we can use th studies as examples in our exploration. However, it is important to recogn that the examples are not necessarily ideal cases; that is, researchers ha made choices that may or may not have been optimal given the questi at hand, as we note. The examples are meant as illustrations of how act research has been conducted, not always as exemplars for future researcl

2.4 Setting Up an Experiment to Test the Effects of a Cause

2.4.1 The Data We Use

The Data Generating Process

We begin our study of causality with the effects of a cause question. We u our example of information and voting as an illustration. We also show h experimental reasoning works within our example. We consider an electi in which there are two or more options before voters who must choose o one. The election might be for President of Mexico, Mayor of Chicago member of the British Parliament, a referendum on a policy proposal, or election created by an experimentalist in a laboratory (where what we me by a laboratory experiment is defined more precisely later). That is, it m be the case that individuals, which we call subjects, have been brought to t laboratory and asked to choose between a set of candidates in an election up by a researcher. The candidates could also be subjects or they might artificial or hypothetical actors. Voters face a choice over whether to vote, a if they vote, which candidate to vote for. We think of the data generated by t election as created by a general data generating process (DGP) that provic the source for the population of data that we draw from in our research.

Definition 2.1 (Data Generating Process or DGP): *The source for the po ulation of data that we draw from in our empirical research.*

The Target Population

The DGP is the source for lots of populations of data, not just one electio When we think of the DGP we think of data generated in all the countries the world (and possibly outside our world). But we are typically interest in just a subset of the data that is generated. What population are interested in? We have to choose a particular target population to study. the election we are studying is a U.S. presidential election, then our targ population includes the data generated by that election. Alternatively, if are conducting an election in a laboratory, then the target population wou

be the population of observations that are generated by such an election set up by the researcher in the laboratory. When we choose to study a particular election or set of elections, we effectively choose a target population. In our analyses, we typically use a sample of data drawn from the target population of data, which is a subset of the target population. The extent that the sample represents the target population is a question of statistical validity and is addressed in Chapter 7.

Definition 2.2 (Target Population): *The population of observations generated by the DGP that an empirical researcher is addressing in his or her analysis.*

2.4.2 What Is an Experiment?

Intervention and Manipulation in the DGP

In an experiment, the researcher intervenes in the DGP by purposely manipulating elements of the environment. A researcher engages in manipulations when he or she varies parts of the DGP so that these parts are no longer naturally occurring (i.e., they are set by the experimenter). We might imagine an experimenter manipulating two chemicals to create a new one that would not naturally occur to investigate what the new chemical might be like. In a laboratory election experiment with two candidates, a researcher might manipulate the information voters have about the candidates to determine how these factors affect their voting decisions. In both cases, instead of nature choosing these values, the experimenter chooses them. Our laboratory election is a particular type of experiment in the social sciences in which subjects are recruited to a common physical location called a laboratory and the subjects engage in behavior under a researcher's direction at that location.

Definition 2.3 (Experiment): *When a researcher intervenes in the DGP by purposely manipulating elements of the DGP.*

Definition 2.4 (Manipulation in Experiments): *When a researcher varies elements of the DGP. For a formal definition of the related concept, manipulated variable, see Definition 3.3.*

Definition 2.5 (Laboratory Experiment): *Where subjects are recruited to a common physical location called a laboratory and the subjects engage in behavior under a researcher's direction at that location.*

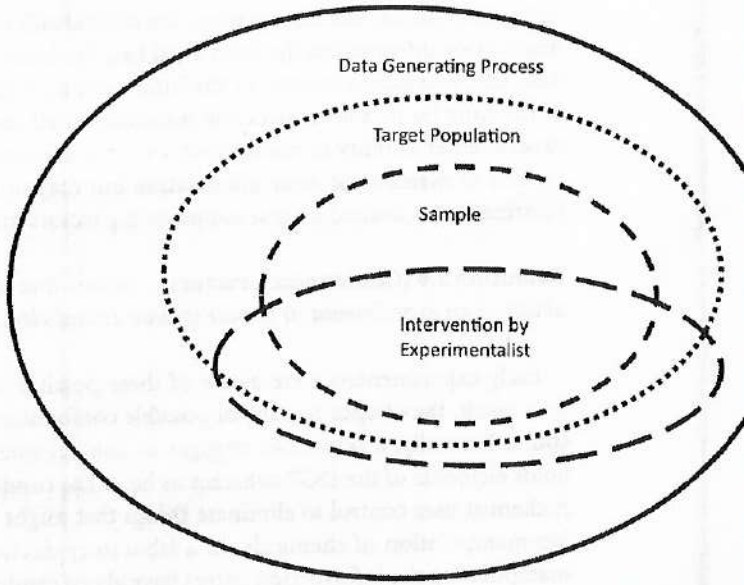

Figure 2.1 Relationships between the Data Generating Process, Target Population, S
ple for Study, and Experimental Intervention.

The intervention and manipulation of the experimenter ideally pri
pally affect the target population in the study (and the sample drawn f
that population that is studied by the researcher). However, the inter
tion and manipulation may also affect other parts of the DGP by affec
choices of individuals outside of the target population. For example, w
a researcher pays subjects for their participation in an experiment, the
ments may affect the income and choices of individuals who are not
of the target population of the experiment as the subjects spend the mo
given to them. Figure 2.1 above illustrates the case in which the interven
affects observations outside the target population (and outside the san
drawn by the experimentalist).

Experimental Control

Confounding Factors. Experimenters worry (or should worry) about fac
that might interfere with their manipulations. For example, trace amou
of other chemicals, dust, or bacteria might interfere with a chemist's ex
iment. That is, the chemist may plan on adding together two chemicals,
when a trace amount of a third chemical is present, his or her manipulat
is not what he or she thinks it is. Similarly, if a researcher is manipulating

information that voters have in a laboratory election, factors such as how the individual receives the information, the individual's educational level, how much prior information the individual has, the individual's cognitive abilities, the individual's interest in the information, or the individual's mood at the time he or she receives the information all may interfere with the experimenter's ability to manipulate a voter's information. The researcher intends to manipulate voter information but may or may not affect voter information as desired if these confounding factors interfere.

Definition 2.6 (Confounding Factors): *Factors that can interfere with the ability of an experimentalist to manipulate desired elements of the DGP.*

Early experimenters were aware of these possible confounding factors. As a result, they began to control possible confounding factors when they could. Formally, a researcher engages in control when he or she fixes or holds elements of the DGP constant as he or she conducts the experiment. A chemist uses control to eliminate things that might interfere with his or her manipulation of chemicals. In a laboratory election, if a researcher is manipulating the information voters have about candidates, the researcher may want to hold constant how voters receive the information and how much other information voters have so that the researcher can focus on the effects of information on how voters choose in the election.

Definition 2.7 (Control in Experiments): *When a researcher fixes or holds constant elements of the DGP to better measure the effects of manipulations of the DGP.*

Observable Versus Unobservable Confounding Factors and the Advantage of the Laboratory. The confounding factors can be of two types: observable and unobservable. Observable factors are simply things that the researcher is able to measure with only random error. For example, in a laboratory election, how the individual receives the information or the individual's educational level are things the researcher can measure arguably with only random error. In contrast, the individual's interest in the information or mood may be something that the researcher cannot observe with confidence. We would call such a factor an unobservable factor. What is observable and unobservable depends on the circumstances of the manipulation and the target population studied. That is, some potential confounding factors such as an individual's educational level may be observable in an experiment conducted with voters participating in a U.S. presidential election as well as in a laboratory election, but it might be easier to observe how much prior

information voters have in a laboratory election than in an experime that is part of a U.S. presidential election. Thus, in the first case, pri information may be observable, but in the latter case, it is unobservable.

As a consequence, to facilitate control, most early experiments in t social sciences, as in the physical sciences, were conducted in laboratori In the laboratory, many confounding factors can be made observable and t experimentalist can then control for their possible interference. As noted the preceding example, in a laboratory election a researcher can, by creati the election that takes place, make observable voters' prior informatio allowing the researcher to better control voters' prior information, whi may be unobservable outside of the laboratory.

Definition 2.8 (Observable Confounding Factors): *Confounding facto that a researcher is able to measure in the target population with only rando error given the experimental manipulation.*

Definition 2.9 (Unobservable Confounding Factors): *Confounding facto that a researcher cannot measure with any confidence in the target populati given the experimental manipulation.*

Baselines in Experiments. One method of controlling confounding variab is to compare experimental results to outcomes in which manipulatio do not occur but all other observable conditions are identical. That is, all the other conditions are held constant – are identical – and the or difference between the two outcomes (the outcome when the manipulati did not occur and the outcome when the manipulation did occur) is t experimental manipulation, then the researcher can argue that the effe he or she is measuring is truly causal; that is, the manipulation has caus any differences between the two outcomes. Oftentimes experimentalists c the outcome in which a manipulation did not occur the "control" a the experiment a "controlled experiment." However, because control is more than just a comparison, but involves other ways that experimentali attempt to control confounding variables, we label such a comparison "baseline comparison." Also, the word control is used in observatio studies in the same general sense: as a method of holding constant the effe of possible confounding variables. We discuss baselines more expansive in Section 8.3.

Definition 2.10 (Baseline): *A manipulation in an experiment designated a researcher as being particularly relevant for comparisons. For a formal de nition of the related concept, baseline treatment, see Definition 8.7.*

Random Assignment

Moving Out of the Laboratory. As Shadish et al. (2002; hereafter SCC) observe, when experimentation moved out of the laboratory and expanded to disciplines such as agriculture, public health, education, and so forth, researchers were no longer able to control adequately aspects of the DGP that might interfere with their manipulations. Researchers could not always find observations with identical observables, and they encountered more unobservable possible confounding factors. A field experiment using human subjects – the name probably comes from agricultural use – is a researcher's intervention that takes place in subjects' natural environments and the researcher has only limited control beyond the intervention conducted.

Definition 2.11 (Field Experiment): *Where a researcher's intervention takes place in subjects' natural environments and the researcher has only limited control beyond the intervention conducted. Usually the relationship between the researcher and the subject is conducted through variables outside of the researcher's control.*

In field experiments in agriculture, it was difficult to use control to account for differences in soil, farming abilities, and so on. It was not possible to find two fields with exactly the same observable conditions and unlikely that unobservable conditions were the same, and researchers expected that these factors could confound the manipulations. Thus, when comparisons were made between the baseline outcome that resulted from no manipulation and outcome that occurred as a consequence of a manipulation, the experimenter could not be sure if the difference was due to the manipulation or to the differences in the fields that he or she could not control.

In field experiments in public health and education, researchers similarly lost the ability to control for variables that could confound their manipulations. It was not possible for them to compare individuals who were exactly the same, living in exactly the same environment, eating the same food, with the same prior health conditions, psychological makeup, or cognitive abilities. Similarly, if a researcher wished to manipulate the information voters have in an election that is naturally occurring as part of the DGP, then the researcher no longer has as much control over how the voters receive information and how much other information voters have in the same way as the researcher can control the information in the laboratory. That is, suppose the information is provided through a mailing about candidate positions on issues. Some voters may not receive the mailing because of mistakes in addresses, others may not check their mail, and still others may

throw the mailing away without reading it. Furthermore, some voters already know the information. These disconnects would occur to a m smaller extent in the laboratory.

As a result of the inability to control factors outside the labora and the difficulty in comparing human subjects, researchers in agricult biomedicine, and social sciences began to develop techniques such as dom assignment as substitutes. Random assignment is when the resear uses a randomization mechanism to assign subjects to manipulations, of which might be a baseline manipulation. In our simple example, researcher may randomly assign some subjects to receive the informa manipulation about the candidates and others to receive no informa (a baseline).

Definition 2.12 (Random Assignment): *When a researcher uses a rand ized mechanism to assign subjects to particular manipulations in the ex ment to better measure the effects of manipulations of the DGP.*

Is Random Assignment Essential for an Experiment? As shown in tion 5.2.2, random assignment can facilitate the ability of researche establish causal inferences. Essentially, because the information is rando assigned across subjects, then the factors that might interfere with the ef of the manipulation, such as whether the subjects actually received the in mation or already knew the information, are in expectation mitigated effects do not disappear, but on average are controlled assuming the domization is effective). The importance of random assignment for ex iments conducted outside of the laboratory in public health, educat and similar disciplines led some to restrict the definition of an exp ment using human subjects to one in which random assignment is u For example, SCC define an experiment explicitly as an intervention uses random assignment, and an intervention that does not is defined quasi-experiment because their focus is largely on experiments condu in these disciplines. Many political scientists have adopted the same con tion. Certainly SCC are correct to say, as we explore later in this book, when one compares two experiments conducted outside the laboratory are exactly alike except that manipulations in one experiment are assig randomly and in the other they are not, the one in which the manipulat are assigned randomly is likely to do better in establishing causal infere than the other, and can certainly do no worse.

However, if we were to compare a laboratory experiment that did not random assignment but the researcher engaged in significant control the elements of the DGP to an experiment outside the laboratory in wh

little control is exercised but random assignment is used, the answer is not so clear. Any experiment with random assignment does not always make "better" causal inferences than any experiment without random assignment. Why? There are two reasons. First, control also facilitates causal inferences, as we discuss in Section 4.1.1. For example, in some laboratory experiments, researchers use what is called a within-subjects design (defined and discussed in Section 3.3.3), which can have advantages over simple random assignment in establishing causality because the same subjects experience all manipulations even if everything else about an experiment is held constant. Subjects serve as their own baselines. Random assignment implies that subjects in expectation have the same probability of experiencing a manipulation, but a within-subject design makes that probability equal to 1 for both manipulations across all subjects.

Second, operationalizing random assignment in experiments is not simple and involves a number of decisions, about what to randomize, across what groups of subjects, and so on, that can affect the value of the inferences made through random assignment. Furthermore, when researchers conduct experiments, especially when the experiments are conducted in the field, issues of response and compliance become important. Nonresponse is when a subject's choices, given manipulations, are not observable, and noncompliance occurs when a subject fails to comply with the manipulation given by the researcher. Random assignment, particularly in field experiments, is thus rarely as ideal for establishing causal inferences as the statistical theory that underlies it would suggest. Thus, both control and random assignments are methods used to deal with factors that can interfere with manipulations; neither is perfect, but both are extremely powerful.

Definition 2.13 (Nonresponse): *Nonresponse is when a subject's choices, given manipulations, are not observable.*

Definition 2.14 (Noncompliance): *Noncompliance occurs when a subject fails to comply with the manipulation given by the researcher.*

Consider some well-known deliberative polling experiments (see Fishkin, 1991, 1993 to 1997; Luskin et al. 2002). In these experiments, a random sample of subjects was recruited to participate in an event to discuss and deliberate public policy on a particular issue. The early experiments suffered from the lack of an explicit baseline sample, noncompliance when subjects who were selected to attend did not, and nonresponse when subjects who attended did not respond to surveys after the event. As a result, many have

argued that these events are not experiments, labeling them quas
experiments, as in the discussion by Karpowitz and Mendelberg (fortl
coming). We agree that the methodological concerns of the critics are ju
tified. The design of the deliberative polls makes it difficult to draw caus
inferences about the effects of deliberation on public opinion. Howeve
not all of these experiments lacked a baseline group and an attempt at ra
dom assignment. For example, Barabas (2004) reports on a deliberative po
event in which a baseline group was surveyed and random samples of sul
jects were recruited to both a baseline group and a group that participate
in the deliberative poll. However, the random assignment was problen
atic because some subjects (fewer than 10%) were recruited to participa
independently by interest groups, some subjects chose not to participa
(noncompliance), and others did not respond when surveyed post-po
(nonresponse). Barabas labels the experiment a quasi-experiment as a co
sequence of these problems with the attempt at random assignment despi
the efforts of the researchers to draw random samples for both the bas
line and manipulated groups. Since almost all field experiments suffer fro
similar problems in implementing random assignment, it would seem th
a strict interpretation of what is an experiment along these lines would ul
mately mean that only a few real field experiments exist in political scienc

Some important and useful experiments have been conducted that d
not use random assignment or baselines or that fail to fully impleme
random assignment, yet they have added significantly to our understandi
of political behavior and institutions just as many experiments in whi
the researcher has little control over variables not manipulated also ha
provided useful knowledge. The fact that a study does not include ra
domization or baselines or the randomization suffers from problems,
our view, does not make it less of an experiment, just as an experime
in which control is minimal is not less than an experiment. As we explain
Section 8.2.4, we think it is important not to confound definitions of expe
ments with normative views of desirable properties because what is desirab
in an experiment depends on the research goal – what the researcher see
to learn – as well as the opportunities before the researcher. What is ide
in an experiment also depends on where it is conducted. In field exper
ments, random assignment can be extremely valuable, although difficu
because control is less available; in the laboratory, the opposite relationsh
holds although both control and random assignment can be much easier
implement. It would be unreasonable for us to define interventions outsi
the laboratory, where there are disconnects between manipulations, ar
what happens to subjects because of a lack of control or problems with th

implementation of random assignment as not really experiments, just as we think it is unreasonable to define interventions without random assignment and baselines as not really experiments. Thus, we define experiments broadly following the traditional definition: an experiment is simply an intervention by a researcher into the DGP through manipulation of elements of the DGP.[8] We further define control and random assignment with or without baselines as usual and important tools by which a researcher can more fruitfully make causal inferences based on his or her interventions. But we recognize that both control and random assignment are rarely implemented perfectly, especially when the experiment is conducted in the field, and thus defining an experiment by whether it contains either one is not useful.

2.4.3 Examples of Information and Voting Experiments

In the Appendix to this chapter contains seven examples of experiments on the relationship between information and political choices. In some cases, the subjects voted in an election conducted by the experimenters; in other cases, subjects reported on their preferences over choices that were presented to them as candidates or choices, sometimes in a hypothetical election, sometimes in an upcoming naturally occurring election in which the subjects would vote or had voted. In some cases turnout decisions, rather than choices between candidates or parties, were measured or surveyed. In all the examples, the experimenters manipulated or attempted to manipulate the information the subjects possessed about the choices before them, how the information was presented to the subjects, or both. All seven used some form of random assignment to manipulations and comparison of manipulations.

The examples, however, illustrate the wide variety of experimental approaches used in political science. Three of the example experiments were conducted during elections in the field. Example 2.1 presents an experiment by Gerber et al. (2007) in which they provided subjects with free newspaper subscriptions during a Virginia gubernatorial election; Example 2.2 concerns an experiment by Wantchekon (2003) during a presidential election in Benin in which he manipulated the campaign messages used by some of

[8] Our definition of an experiment is also traditional in that researchers used experimentation for many years before the advent of random assignment as a tool in establishing causal inferences in the early twentieth century. If we label interventions into the DGP as nonexperiments if they do not use random assignment, many famous and infamous research trials would be considered nonexperiments, such as Edward Jenner's research leading to the smallpox vaccine and the Tuskegee syphilis study discussed in Section 11.4.1.

the political parties; and Example 2.3 discusses an experiment by Clir and Lapinski (2004) during the 2000 U.S. presidential election, in wl Clinton and Lapinski showed subjects negative campaign advertiseme Clinton and Lapinski's experiment is an Internet survey experiment bec it is embedded in a survey and conducted via the Internet. We discuss tl particular types of experiments more expansively in Section 8.2.1.

The other four examples are laboratory experiments. However, they across important dimensions. Two were conducted by political psycholo involving hypothetical candidates (Example 2.4, experiments conducte Kulisheck and Mondak (1996), Canache et al. (2000), and Mondak Huckfeldt (2006), and Example 2.5, conducted by Mutz (2007)). Mon and his coauthors varied how much information subjects had about ca date qualities, whereas Mutz varied the visual mechanism by which subj learned about candidates. The remaining two examples were conducte political economists in which subjects chose in a laboratory election were given payments based on which choices won (Example 2.6, condu by Battaglini, Morton, and Palfrey, and Example 2.7, conducted by Dasg and Williams [2002]). In Dasgupta and Williams's experiment, the cho before voters were called candidates, and voters were told they were vo in an election, whereas in Battaglini, Morton, and Palfrey's (2008, 2(study, subjects were asked to guess the color of an unseen jar given in mation provided to them and the "winner" was the color that received majority of guesses. Some of the experiments were conducted via comp networks. (Battaglini et al., Dasgupta and Williams, and Mondak et al. u computers in a laboratory; Clinton and Lapinski's experiment was c ducted via the Internet). Mondak et al. and Mutz also used other meth to measure how subjects used or responded to information: Mondak e measured the time taken by subjects to respond to various questions Mutz measured skin reactions to visual presentations of information.

We can also see tremendous variation in the control used in the exp ments. In Examples 2.1, 2.3, and 2.5, Gerber et al., Clinton and Lapinski, Mutz, respectively, designate one of the manipulations as a baseline ma ulation. In other examples, although comparisons are made, the ma ulation that one would call a baseline is not so obvious, as in Dasg and Williams's and Wantchekon's experiments. In the field experime the researchers generally had little control over many possible confoun variables. For instance, in Example 2.1, Gerber, Kaplan, and Bergan h little control over what is reported in the newspapers about the elect whether subjects read the newspaper articles about the election, and to s extent whether their access to the newspaper is manipulated. In cont in the laboratory experiments in Example 2.5, the subjects came to Mu

laboratory and watched videos she had prepared in a controlled setting. In studies by Dasgupta and Williams and Battaglini et al., the researchers used financial incentives in an attempt to control subjects' preferences over their choices in the elections, which was not possible in the other five examples. That is, in the other five examples, the researchers must account for partisan preferences held by the subjects that may also affect their choices but cannot control them explicitly.

The seven examples also differ in the types of subjects used. The field experiments used residents in the localities where they were conducted, whereas the laboratory experiments generally used students, with the exception of Mutz's experiment which used nonstudents recruited via employment agencies or civic groups. Clinton and Lapinski used a national sample from an Internet survey organization. Some of the laboratory experiments that used students drew subjects from multiple universities: Mondak et al. used students from the United States, Mexico, and Venezuela, and Battaglini et al. used students from Princeton University and New York University. We discuss the advantages and disadvantages of different types of subjects in Chapter 9.

Finally, all seven examples reference to varying degrees one or more of the theories of voting mentioned in Section 2.3.3. Gerber et al., Mondak et al., and Mutz reference either explicitly or implicitly Cognitive Miser views of voting in which information primes, frames, or persuades voters. Clinton and Lapinski contend that their experiment provides important new evidence on negative advertisements, while Wantchekon argues that his work similarly informs our understanding of how clientelism works. Dasgupta and Williams and Battaglini et al. relate their experimental work to pivotal voter models. And both Gerber et al. and Battaglini et al. discuss expressive theories of voting. We address these and other variations in the examples (and additional examples presented later) throughout the text.

2.4.4 What Is Not an Experiment?

Qualitative Research and Traditional Surveys

Although our definition of experiments is encompassing, it excludes other research with human subjects in the social sciences such as interviews and qualitative, soak-and-poke, political science research that aims not to intervene in the DGP but to measure and observe how the DGP operates through close interaction with human subjects. Manipulations that occur in these types of research studies are not generally purposeful, but accidental. Whether, of course, it is possible to observe in close interaction without manipulating the DGP is an important issue in such research, but the overall

goal, as we understand it, is access to human subjects in the DGP as if t researcher were not there, rather than to manipulate the DGP. If qualitati researchers see themselves as intervening with the purpose of altering t DGP, we call that research an experiment. Similarly, a traditional surv is not an experiment because the goal of the researcher is to measure t opinion of the respondents, not to intervene or manipulate elements of t DGP that affect these opinions. When a researcher does purposely attem to use a survey to manipulate elements of the DGP that theoretically affe respondents' opinions, we call this an experiment. We discuss experimer in surveys more expansively in Section 8.2.1.

Note that we recognize that the goal of many experimental manipulatio is to better measure political behavior or preferences, as in many of t political psychology experiments which use implicit messages in an attem to better measure racial prejudices (see, for example, Lodge and Tab 2005). Yet the means of achieving the goal is through manipulation, n passive observation, which makes this research experimental rather th observational, in our view.

Natural and Policy Experiments and Downstream Benefits of Experimen

Sometimes nature acts in a way that is close to how a researcher, given t choice, would have intervened. For example, hurricane Katrina displac thousands of New Orleans residents and changed the political makeup of t city, as well as having an impact on locations that received large numbe of refugees. Although no political scientists we know would wish such disaster to occur in a major city, many find the idea of investigating t consequences of such a manipulation an exciting opportunity to evalua theories of how representatives respond to changing constituencies, f example. Katrina was an act of nature that was close to what a politic scientist would have liked to have done if he or she could – interveni and changing the political makeup of several large U.S. cities such as Ne Orleans, Houston, and Atlanta.

Natural manipulations might also occur in our information and voti example. For instance, in the case where the mailing described earlier in naturally occurring election is provided without input by a researcher, th it is a natural manipulation. When natural manipulations occur, sometim researchers argue that the manipulation is "as if" an experimentalist mani ulated the variable. The researcher often calls the manipulation a "natu experiment," although the name is an oxymoron because by definition a experiment cannot be a situation where the DGP acts alone and thus we c not call these experiments, according to our definition. The researcher contending that nature has two sides: the side that generates most data, ar

then the interventionist side that occasionally runs experiments like academics, messing up its own data generating process. Even though in this case the researcher is not doing the intervening, the approach taken with the data is as if the researcher has. When does it make sense for a researcher to make such a claim and approach his or her observational data in this fashion? The answer to this question is complicated and we address it fully in Section 5.4.3. Example 2.8 in the Appendix presents a study of a natural experiment involving the effect of information on voter turnout by Lassen (2005).

Definition 2.15 (Natural Experiment): *Nonexperimental or observational data generated by acts of nature that are close to the types of interventions or manipulations that an experimentalist would choose if he or she could.*

A special type of natural experiment occurs when government officials manipulate policies. For example, in Lassen's experiment, the natural manipulation occurred when government officials varied the ways in which public services were provided. Similarly, De La O (2008) exploits governmental policy changes in Mexico to consider how different governmental services impact voter participation. We call such a manipulation a policy experiment when it is undertaken by government officials without academic involvement.

Definition 2.16 (Policy Experiment): *A field experiment in which a government agency or other institution chooses to intervene and act "like an experimentalist."*

More recently, governments and other nonacademic institutions have formed collaborations with academic researchers to conduct experimental manipulations. An example of such a collaboration is provided in Example 12.1, where Olken (2008) collaborated with officials to manipulate the mechanisms by which villages in Indonesia made decisions about which public projects to fund. When such collaboration occurs and the researcher is directly involved in consciously choosing the design of the manipulation, the manipulation is an experiment, as we have defined. If the collaboration does not involve a researcher in the design process but simply allows a researcher the opportunity to gather data on a manipulation already planned and designed for a different purpose, the research is a natural experiment and not an experiment, as we have defined.

Occasionally researchers might use the manipulation of an experiment conducted in the past to investigate either a new hypothesis or the

long-term implications of the original manipulation. Gerber and Gre (2002) label such research downstream research and the results of th investigations the downstream benefits of experimentation. The use of p vious experiments in this fashion certainly has advantages in the same w that natural and policy experiments can be useful in empirical analysis. Sondheimer (forthcoming) for a discussion of how researchers can ben from prior manipulations.

Definition 2.17 (Downstream Benefits): *The benefits of analysis of pr experiments, either conducted by academics for research or conducted by g ernment as a policy experiment.*

Computational Models and Simulations

Occasionally political scientists who use computational or agent-ba models to numerically solve formal models of politics call the output "exp iments," or others who run counterfactual analyses using parameters e mated from an empirical model call their analyses "experiments" (see, e Kollman and Page, 1992). Computer simulations to solve formal models aids in solving a model, not in "testing" the model with empirical data. Th simulations are an extension of the researcher's brain. Similarly, counterf tual simulations using parameters from estimated empirical models are a to understanding the empirical model estimated using either observatio or experimental data and are an extension of the analysis of those da not the generation of new experimental data. In experiments the subje make "real" decisions and choices and are independent of the researcl and new data are generated. The subjects are not simulating their behav but engaging in behavior. The environment created by the experimenta is not an observational environment but it is real in the sense that individuals are involved. Thus, simulations and experiments serve enti distinctive purposes.[9]

Counterfactual Thought Experiments

Related to the use of computational or agent-based models to solve for models are what have been called counterfactual thought experiments, which nonformal theorists hypothesize the effects of situations in wh

[9] Some experimentalists believe that this also means that it is important that the exp mentalist not "hard-wire" the experiments by telling subjects how to choose or beh However, there is no hard-and-fast rule on such suggestions, because in some cases do so may be an important part of the experiment and is the subject of the experime investigation.

one or more observables takes on values contrary to those observed (see, e.g., Tetlock and Belkin [1996] and Tetlock et al. [2006]). For example, what would have happened if Great Britain had confronted Hitler more before World War II? Again, these are not experiments as we have defined them because they are extensions of the researcher's brain – theoretical speculation about what would have happened historically if a variable had been manipulated.

2.4.5 Experimental Versus Observational Data

In most experimental research, the variation in the data is partly a consequence of the researcher's decisions before the data are drawn and measured. If we think of the DGP before or without intervention as nature acting alone, then the DGP after intervention is nature and the researcher interacting. We call the data generated by such intervention "experimental data." So for instance, in Clinton and Lapinski's experiment, Example 2.3, the 2000 presidential election without the experiment would be nature acting alone, but with the intervention of the researchers is nature and Clinton and Lapinski interacting. Data on the presidential election that do not involve such interaction are observational data (such as who the candidates were), but data generated through the interaction (such as how the subjects voted after having the campaign advertisements they saw manipulated by the researchers) are experimental data.

Definition 2.18 (Experimental Data): *Data generated by nature and the intervention of an experimentalist.*

Nonexperimental empirical research involves using only data drawn from the population in which all variation is a consequence of factors outside of the control of the researcher; the researcher only observes the subset of data he or she draws from the DGP but does not intervene in that process or if he or she does so, it is an accidental intervention, not purposeful. There are many observational studies of the 2000 U.S. presidential elections of this sort. This approach assumes, of course, that the researcher can and will choose to measure the data perfectly; clearly, choices made in measurement can result in a type of post-DGP intervention, but the data are still not experimental data because the data are generated without intervention.

Some distinguish between experimental and "naturally occurring" data. Others talk of the "real world" versus the experimental world. Such terms are misleading because nature is also involved in determining variation in

experimental data. Even in laboratory experiments, although the experimenter may intervene in the data generating process, the subjects in the experiment who make decisions are "real" and their decisions occur "naturally," albeit influenced by the experimental environment. Since as political scientists we are interested in human behavior, we should recognize that the humans participating in an experiment are as "real" as the humans not participating in an experiment.[10] A more neutral description of research using only data where the variation is a consequence of factors outside of the control or intervention of the researcher is research that uses observational or nonexperimental data; we use that terminology.

Definition 2.19 (Nonexperimental or Observational Data): *Data generated by nature without intervention from an experimentalist.*

2.5 Chapter Summary

Fundamentally, scientific research is about building and evaluating theories about the causes of effects. Political scientists are interested in studying why and how people vote, for example. One of the ways we build toward such theories and evaluate them is to study the effects of causes. Many theories of why and how people vote make causal predictions about how information, a cause (either in content or presentation), affects voters' choices, an effect. In this chapter we have reviewed some of these theories of voting and their predictions about the relationship between information and voting as an illustration. To evaluate theoretical predictions, or sometimes just hunches about the relationship between information and voting, many political scientists have used experiments. We have discussed examples of these in this chapter and the examples are presented more fully in the Appendix.

In this chapter we have also surveyed the features of the experimental method used by researchers to address predictions. In summary, the standard use of the experimental method in political science typically involves the following four principal features:

1. Designating a target population for the experimental study,
2. Intervention and manipulation in the DGP,

[10] We discuss in Chapter 7 whether the experimental environment leads subjects to make choices they would not make if the same changes in their environment would occur via nature or the DGP, rather than through experimental intervention. Even if "being in an experiment" has such an effect, this does not mean that the subjects' choices are less real or less human. It means that we must understand that sometimes the aspects of the experiment itself that cause such effects are treatments that must be considered explicitly

7

Validity and Experimental Manipulations

In the previous chapters we have examined both experimental and nonexperimental research, taking a largely common perspective. Although we have mentioned some of the differences between experimental work and nonexperimental analysis and some of the different types of experimental research, we have generally focused on commonalities rather than distinctions. Yet, the differences in approaches can be important and many are controversial. In this part of the book we turn to these differences. Usually the controversies have to do with arguments about the validity, robustness, or generality of particular experimental designs. Thus, before we turn to the specific differences, we begin with a review of the concept of validity in research. Then we turn to particular and sometimes controversial issues in experimentation such as the location of an experiment (whether lab or field), the subjects recruited (whether subjects are students), and how the subjects are motivated (whether financial incentives are used).

7.1 Validity of Experimental Research

Suppose that we have conducted some empirical research with either experimental or nonexperimental data. We ideally want a research design that will provide us with *valid* results that are true for the population we are analyzing and *robust* results that *generalize* beyond our target population (see Definition 2.2).[1] So, for example, before we begin to study the effect of information on voting – the effect of changes in T_i on Y_i – we would like to come up with an experimental design that will give us results that meet these criteria.

[1] We use the term population here rather than data set because questions about how the data set relates to the population are some of the issues in establishing internal validity, which is discussed later.

7

Validity and Experimental Manipulations

In the previous chapters we have examined both experimental and nonexperimental research, taking a largely common perspective. Although we have mentioned some of the differences between experimental work and nonexperimental analysis and some of the different types of experimental research, we have generally focused on commonalities rather than distinctions. Yet, the differences in approaches can be important and many are controversial. In this part of the book we turn to these differences. Usually the controversies have to do with arguments about the validity, robustness, or generality of particular experimental designs. Thus, before we turn to the specific differences, we begin with a review of the concept of validity in research. Then we turn to particular and sometimes controversial issues in experimentation such as the location of an experiment (whether lab or field), the subjects recruited (whether subjects are students), and how the subjects are motivated (whether financial incentives are used).

7.1 Validity of Experimental Research

Suppose that we have conducted some empirical research with either experimental or nonexperimental data. We ideally want a research design that will provide us with *valid* results that are true for the population we are analyzing and *robust* results that *generalize* beyond our target population (see Definition 2.2).[1] So, for example, before we begin to study the effect of information on voting – the effect of changes in T_i on Y_i – we would like to come up with an experimental design that will give us results that meet these criteria.

[1] We use the term population here rather than data set because questions about how the data set relates to the population are some of the issues in establishing internal validity, which is discussed later.

Although political science has become more experimental, the most controversial questions raised about experimental research have to do with the validity, robustness, or generalizability of that research for answering substantive questions in political science. Can an experimental study of the effect of information on voting give us results that are valid and robust? We spend most of this book addressing questions of validity and robustness of all empirical research. But first we need to define these terms more precisely and deal with some of the confusions in the literature. The first such confusion is over the definitions of what we mean by validity and the types of validity. The essence of the validity of empirical research is the question: "What can we believe about what we learn from the data?" Shadish et al. (2002) or SCC (recall Section 2.4.2) use the term validity as the "approximate truth" of the inference or knowledge claim. So suppose we conduct a study, either experimental or observational, of the relationship between information and voting and we ask the validity question: What do these data tell us? This definition, however, leaves unanswered the population over which one establishes the approximate truth. We have defined the data generating process (DGP) as the source for the population from which we draw the data we use in empirical research. However, usually when we engage in empirical research we consider only a subset of the population of data generated by the DGP. For example, we may want to explain voter turnout in the United States. We probably would not be interested in the data on turnout in China in such a study.

Definition 7.1 (Validity): *The approximate truth of the inference or knowledge claim.*

When we think of validity, do we mean valid with respect to the target population of the research or is it another different population of observations? Such questions have typically been divided into two separate validity issues. This simplistic view of how to refine the concept of validity is based on the early division of Campbell (1957) and is universally used by political scientists. Specifically, political scientists generally use *internal validity* to refer to how valid results are within a target population and *external validity* to refer to how valid results are for observations not part of the target population.[2] So if our data, for example, are drawn from a U.S. election,

[2] Other terms have been used by researchers to capture the issues that we address about validity in this chapter. For example, Levitt and List (2007b) use *generalizability* and others use *parallelism* as in Wilde (1981) and Smith (1982).

the internal validity question would ask how valid our results are from the analysis of the data for the target population of voters in that U.S. election. The external validity question would ask how valid our results are for other populations of voters in other elections, in the United States, elsewhere in the world, or in a laboratory election.

Definition 7.2 (Internal Validity): *The approximate truth of the inference or knowledge claim within a target population studied.*

Definition 7.3 (External Validity): *The approximate truth of the inference or knowledge claim for observations beyond the target population studied.*

However, this simplistic division of validity masks the complex questions involved in establishing validity and the interconnectedness between internal and external validity. Both internal and external validity are multifaceted concepts. In this chapter we explore both types of validity.

7.2 Deconstructing Internal Validity

As SCC and McGraw and Hoekstra (1994) discuss, nearly 40 years ago Campbell abandoned the simple binary division of validity into internal and external that still dominates political science. It is ironic that political scientists typically cite him as the authority when they use the simplistic terms. Cook and Campbell (1979) extended validity into a typology of four concepts. SCC use this typology by incorporating clarifications suggested by Cronbach (1982). In this typology, validity is divided into four types: construct, causal, statistical, and external.[3] The first three of these types together are what political scientists think of as internal validity. By exploring

[3] Cook and Campbell (1979) called causal validity "local molar causal validity." SCC explain how the term local molar causal validity explains itself (2002, p. 54): "The word *causal* in *local molar causal validity* emphasizes that internal validity is about causal inferences, not about other types of inferences that social scientists make. The word *local* emphasizes that causal conclusions are limited to the context of the particular treatments, outcomes, times, settings, and persons studied. The word *molar* recognizes that experiments test treatments that are a complex package consisting of many components, all of which are tested as a whole within the treatment condition." SCC label local molar causal validity "internal validity" because they believe that the longer term is too unwieldy and that this is what Campbell originally viewed as internal validity. Given that many political scientists think of internal validity as whatever is left over after external validity and thus includes statistical, causal, and construct validity, we define internal validity differently from SCC. SCC also call statistical validity "statistical conclusion validity." We use the shorthand terms causal and statistical validity because they are easy to remember and capture the essence of these types of internal validity.

how each type represents a distinct question, we can better understand the different challenges involved in determining internal validity. How empirical research, either experimental or observational, establishes the validity of two of these types, causal and construct, was the focus of the previous four chapters. But before turning to these types of validity, we address statistical validity.

Definition 7.4 (Construct Validity): *Whether the inferences from the data are valid for the theory (or constructs) the researcher is evaluating in a theory testing experiment.*

Definition 7.5 (Causal Validity): *Whether the relationships the researcher finds within the target population analyzed are causal.*

Definition 7.6 (Statistical Validity): *Whether there is a statistically significant covariance between the variables the researcher is interested in and whether the relationship is sizable.*

7.2.1 Statistical Validity

Problems of Statistical Validity

Statistical validity is defined as whether there is a statistically significant covariance between the variables the researcher is interested in and whether the relationship is sizable. Suppose we find a relationship between information and voting (i.e., between T_i and Y_i) as defined in Chapter 3; statistical validity is whether the relationship is significant and sizable. Essentially this is what is often called the estimation problem of statistical analysis. Given the assumptions the researcher has made about the variables studied in the given data set, are the estimates efficient, accurate, significant, and sizable? Is the data set representative of the target population? Although these concerns seem minor compared to other matters we address later, as any empirical researcher knows, estimation is not an open-and-shut case. What does it mean when a researcher finds that the statistical relationship is just on the edge of the standard level of significance of 5%? Many now advocate an approach that focuses on reporting the actual significance level rather than just whether a result passes a threshold. Another question involved in statistical validity is whether the statistical assumptions about the distributions of the variables are supported. Are the errors estimated correctly? Is the size of the relationship consequential or not? How do we evaluate the significance of interaction terms?

Estimation issues can be important and are sometimes overlooked. As we discuss in Section 5.6, a popular method of empirical research which springs from experimental reasoning is what has been called the "difference-in-differences" approach to studying the effects of state laws or state policies. Researchers compare the difference in outcomes after and before the law or the new state policy for those affected by the law or state policy to the same difference in outcomes by those who have not been affected by the law or new state policy. The researchers often use ordinary least squares (OLS) in repeated cross sections or a panel of data on individuals before and after the passage of the law or new state policy. They then use the coefficient estimated for the dummy variable in the OLS that represents whether the law applies to the given observation as an estimate of the effects of the law or policy. However, Bertrand et al. (2004) pointed out that the OLS estimations are likely to suffer from possible severe serial correlation problems which when uncorrected lead to an underestimation of the standard error in estimating the coefficient and a tendency to reject null hypotheses that the law or policy has no effect when the null hypothesis should not be rejected.

The serial correlation occurs for three reasons: (1) the researchers tend to use fairly long time series, (2) the dependent variables are typically highly positively serially correlated, and (3) the dummy variable for the existence of the law or policy changes vary little over the time period estimated. The authors propose a solution – removing the time-series dimension by dividing the data into pre- and post-intervention periods and then adjusting the standard errors for the smaller number of observations this implies. They also point out that when the number of cases is large – for example, if all 50 states are included – then the estimation is less problematic. This is just one example of how statistical validity can matter in determining whether results are valid.

Statistical Replication

Statistical replication is a powerful method of verifying the statistical validity of a study. We follow Hunter (2001) and Hamermesh (2007) in dividing replication into two types. *Statistical replication* is when a researcher uses a different sample from the same population to evaluate the same theoretical implications as in the previous study or uses the same sample but a different statistical method evaluating the same theoretical implications (which some call verification), in both cases holding the construct validity of the analysis constant. *Scientific replication* is when a researcher uses a different sample, uses a different population to evaluate the same theoretical constructs,

or uses the same sample or a different sample from either the same or different population focusing on different theoretical implications from those constructs. We discuss scientific replication when we address external validity.

Definition 7.7 (Statistical Replication): *When a researcher uses a different sample from the same population to evaluate the same theoretical implications as in the previous study with equivalent construct validity or uses the same sample from the same population but comparing statistical techniques to evaluate the same theoretical implications as in the previous study, again with equivalent construct validity.*

It is easy to see that statistical replication is concerned with statistical validity rather than the external validity of results. In fact, researchers working with large data sets would probably be well served to engage in cross-validation, where the researcher splits the data into N mutually exclusive, randomly chosen subsets of approximately equal size and estimates the model on each possible group of $N-1$ subsets and assesses the model's predictive accuracy based on each left out set. Although statistical replication may seem mundane, Hamermesh presents a number of interesting situations in economics where statistical replication has led to controversy.

There are examples in political science where results have been verified and called into challenge. For instance, Altman and McDonald (2003) showed that variations in how software programs make computations can, in sophisticated data analysis, lead to different empirical results in a statistical replication. In political science, statistical replication with new samples from the same target population can also lead to different results and some controversy. For example, Green et al. (1998) replicated analyses of MacKuen et al. (1989, 1992) on macropartisanship using a larger data set from the same population, calling into question the original conclusions of the analysis.[4] Because of the possibility that statistical replication may lead to different results, many political science journals now require that authors make their data plus any other necessary information for replicating the analysis available to those who may be interested. There are, of course, a number of issues having to do with the confidentiality of different data sets and sources; nevertheless, the general perspective within political science is that efforts should be made to make replication of statistical analysis possible.

[4] See also the response by Erikson et al. (1998).

In terms of experimental work, replication can at times be a bit more complicated, unless it is the simple verification variety as in Imai's (2005) statistical replication of Gerber and Green's (2000) mobilization study.[5] Statistical replication that involves drawing a new sample from the same population requires that a new experiment be conducted using subjects from the same target population with the same experimental protocols. Oftentimes experimentalists do this as part of their research, conducting several independent sessions of an experiment using different samples of subjects from the same pool.

7.2.2 Causal Validity and the Identification Problem

Even if the results are statistically valid, if we want to be able to say something about the effects of causes, then we need for our results to have causal validity. Causal validity is the determination of whether the relationships the researcher finds within the target population analyzed are causal. Thus, suppose we find a relationship between information and voting behavior in an election, either observational or experimental. Establishing causal validity for that relationship would mean establishing that changes in one of the variables – we posit information – causes changes in the other variable – voting behavior. Formally, using the preceding notation, changes in T_i cause changes in Y_i. We have spent the previous four chapters exploring how a researcher establishes causal validity using either the Rubin Causal Model or the Formal Theory Approach (FTA).

A concept closely related to causal validity is the notion of *identification* in econometrics. As Manski (1995, 2003) explains, econometricians have found it useful to separate out the concerns of identifying relationships from the concerns in estimating relationships. Estimation problems have to do with statistical issues of whether, given the data set analyzed and the assumptions made about the relationship between the data set and the population, the parameters of interest are efficiently and consistently estimated, or statistical validity.[6] Manski remarks (2003, p. 12): "Statistical inference seeks to characterize how sampling variability affects the conclusions that can be drawn from samples of limited size."

[5] See also Gerber and Green's (2005) response.

[6] Consistent parameter estimates are those that, under the assumptions made about the population, converge on the true population parameters as the sample size of the data set analyzed grows without bound. Efficient estimates are loosely those that have the lowest possible variance of unbiased estimators.

In contrast, an identification problem exists when it is problematic to establish causal inferences even if the researcher has an unlimited sample from the population. Identification problems exist in many contexts. Of particular interest to political scientists is the identification problem that occurs because we cannot observe the same individual in multiple states of the world in the DGP. For example, suppose we are interested in the causal effect of education on voting. Our population is the citizens in a particular region. We cannot simultaneously observe each citizen, both educated and uneducated. Even if we have an unlimited sample from the population, we would not be able to find such observations. We can make assumptions about the probability of being educated and the reasonableness of comparing educated citizens' choices with those of uneducated citizens (and in rare cases observe them in the two states sequentially).

This type of identification problem is often labeled a selection problem because in observational analysis individuals select their education levels; they are not manipulated by the researcher. However, the problem is more fundamental than this label suggests. The difficulty arises because counterfactual observations are impossible to observe even if education could be randomly assigned to individuals – we still cannot observe the same individual both educated and uneducated. As we discussed in earlier chapters, there are experimental designs which come close to providing pseudo-counterfactual observations, and random assignment does help one "solve" the problem under particular assumptions. But even these solutions are merely close; they do not fully capture human choices in multiple states of the world simultaneously.

7.2.3 Construct Validity

Defining Construct Validity

When some political scientists think of internal validity, particularly with respect to experiments that evaluate formal models or take place in the laboratory, they are often referring to what SCC call construct validity. Construct validity has to do with how valid the inferences of the data are for the theory (or constructs) the researcher is evaluating. Essentially, establishing construct validity is an essential part of estimating causality in FTA; in FTA the goal is to investigate causal relations within a research design that has construct validity. Thus, if we think about causal validity as establishing whether changes in T_i cause changes in Y_i, construct validity takes a broader look and asks if our empirical analysis is a valid evaluation of our theory or model about why changes in T_i cause changes in Y_i.

In experimental research the question is whether the design of the experiment is such that the variables investigated are closely equivalent to the variables the theory is concerned with. Are those things that the theory holds constant held constant in the experiment? Are the choices before the subjects the same as the choices assumed in the theory? Do the subjects have the same information about each other and about the environment that the theory assumes? Are the subjects in the experiment from the same target population that the theory addresses? In other words, is there a close match between what the theory is about and what is happening in the manipulated DGP?

In observational studies, researchers who work with formal theoretical models think about the equations underlying the theory and the equations underlying the empirical analysis and their relationship. Is the estimated empirical model derived from or equivalent to the underlying theoretical model? If there are disconnects between the empirical model and the theoretical model, to what extent do these disconnects lead one to discard the results of the research as not being relevant to the theory? These are issues that we have already addressed extensively in Chapter 6.

Construct Validity and the Generalizability of Results

Although we group construct validity as part of internal validity, as do most political scientists, doing so misses an important aspect of construct validity that makes it more than just about a given experiment. Construct validity is also about generalization. The generalization is to a theoretical construct that ideally the researcher does not view as limited to the particular empirical analysis, but a more general theory. Because of this, being able to establish construct validity can actually help build answers to external validity questions about the theory and any analysis of the theory. As SCC argue (2002, p. 93):

> [V]alid knowledge of constructs that are involved in a study can shed light on external validity questions, especially if a well-developed theory exists that describes how various constructs and instances are related to each other. Medicine, for example, has well-developed theories for categorizing certain therapies (say, the class of drugs we call chemotherapies for cancer) and for knowing how these therapies affect patients (how they affect blood tests and survival and what their side effects are). Consequently, when a new drug meets the criteria for being called a chemotherapy, we can predict much of its likely performance before actually testing it (e.g., we can say it is likely to cause hair loss and nausea and to increase survival in patients with low tumor burdens but not advanced cases). This knowledge makes the design of new experiments easier by narrowing the scope of pertinent patients and outcomes, and it makes extrapolations about treatment effects likely to be more accurate.

In political science this is also true when a researcher works with a well-developed theory. Results from experiments with high construct validity can help us answer more general questions than those without construct validity. For example, Wittman (1983) and Calvert (1985) demonstrate in a two-candidate model of spatial competition, if the candidates have different policy preferences independent of whether they are elected and there is uncertainty about the ideal point of the median voter in the electorate, the candidates will choose divergent policy platforms in equilibrium. However, if the candidates are certain about the location of the median voter's ideal point, then the candidates converge in equilibrium. This comparative static prediction has been supported in experiments designed to have high construct validity (see Morton, 1993).

The theoretical prediction also has implications for the relationship between factors that affect whether candidates have policy preferences (such as candidate selection mechanisms) and knowledge of voter preferences and the divergence of candidate policy positions. We can extrapolate from the theory to consider other possible relationships for future empirical investigation, such as how a change in a candidate selection mechanism that makes candidates more ideological may impact candidate policy positions. For example, we may argue that in open primaries (where all registered voters are allowed to participate in the selection of candidates) candidates are less ideological than in closed primaries (where only the voters registered in a particular party are allowed to participate in a party's primary). The theory, supported by the experimental results in one target population, would then suggest that there is more divergence between candidates in closed primaries than in open primaries. Indeed, Gerber et al. (1998) find that this new theoretical prediction is supported with data on the policy positions of congressional incumbents, a different target population. Their research shows that congressional incumbents' policy positions are closer to the estimated positions of the median voters in their districts in states with open primaries than in states with closed primaries.

Although the preceding example demonstrates how empirical research with high construct validity that supports a theory in one target population can be useful as a basis for generalizing beyond the initial empirical research to implications in other target populations, a negative result from empirical research with high construct validity can also be useful. If the empirical research shows that the theory's predictions do not hold in one target population, and the research has high construct validity, then the results from the analysis can help develop a more general and robust theory, leading again

to new predictions about other populations beyond the target population in the original empirical analysis.

Construct Validity and External Validity

The previous section argues that construct validity of studies allows for generalization beyond those studies. The quote from SCC suggests that studies with construct validity can shed light on external validity questions. However, we do not believe such a conclusion should be taken too far. In our opinion, construct validity is not a substitute for external validity. To see why this is the case, consider what is implied by results from studies with construct validity. Suppose a researcher finds a situation in which the empirical research is considered to have construct validity and the theory's behavioral predictions are not supported. Does that mean that we should always change the theory once we find a single negative result? Although the empirical study may be considered to have construct validity, it is unlikely that a single negative result would be seen as decisive in determining the merits of the theory. Why? This is because all theories and models are abstractions from the DGP and, therefore, all have parts that are empirically false and can be proven empirically false when confronted with some observations of the DGP.[7] The question is not whether a theory can be proven empirically false, but when empirical inconsistencies with the theory matter enough for the theory to be modified or even discarded.

Similarly, suppose a researcher, again conducting empirical research considered to have construct validity, finds that the theory's behavioral predictions *are* supported. Does that mean that we should unconditionally accept the theory? Not necessarily. In our opinion, theory evaluation in the social sciences is a cumulative process that occurs through replication and complementary studies. However, because any theory can be disproved with enough data, the evaluation of theory is not purely an empirical question. As with Fudenberg (2006), we believe that theory should be judged on Stigler's (1965) three criteria: accuracy of predictions, generality, and tractability. In conclusion, construct validity is a property of a particular empirical study. However, negative or positive results from one such empirical study with construct validity are rarely adequate even if the results are strong and robust enough to accept or reject the theory. In our opinion, as we explain later, to establish external validity of results further empirical

[7] Note that a theory is always true in a theoretical sense if it is logically consistent; that is, the results or predictions follow directly from the assumptions.

study is required, both nonexperimental and experimental if possible, to fully evaluate the value of social science theories.

7.2.4 Summary of Internal Validity

When political scientists refer to internal validity, they are often referring to three distinct and important aspects of validity: statistical, causal, and construct. It is better if we think of these issues separately, because each involves a different type of answer and has a separate set of concerns. It is quite possible – in fact, highly likely given advances in estimation techniques – that an analysis satisfies statistical validity but not causal validity or construct validity, although in some cases advances in the study of identification problems (causal validity) have outpaced estimation procedures, as discussed by Athey and Haile (2002) and exemplified in the problems discussed earlier with difference-in-differences studies.

7.3 Deconstructing External Validity

7.3.1 External, Statistical, and Ecological Validity

In contrast to internal validity, external validity is a more widely understood idea among political scientists – if you asked an average political scientist the definition of external validity, he or she would probably give you something similar to what we have written earlier. But knowing a definition and applying it are not the same, and political scientists often apply the term external validity incorrectly even if they are aware of the definition. For example, sometimes when a political scientist claims that an experiment does not have external validity, he or she is making the claim that the result is not internally valid in a statistical sense – that the sample is not a random sample from the appropriate target population and thus the conclusions are not statistically valid for the appropriate target population. But random sampling from a target population does not mean that a result is externally valid. If a researcher draws a random sample from the U.S. population to evaluate a hypothesis, the results of the analysis are not necessarily externally valid to individuals in China. External validity has to do with generalizing to populations beyond the target population, so whether one has a random sample from the target population tells one nothing about the external validity for other populations for which one has not taken a random sample.

Other times political scientists confuse external validity with *ecological validity*. Ecological validity, however, is not about the validity of results

from empirical research. It is about the similarity between the environment constructed in the research and a target environment. Some experimentalists call this mundane experimental realism or contextual congruence. The experimental environment is considered ecologically valid if the methods, materials, and settings of the research are similar to the target environment. Ecological validity is similar to what Harrison and List (2004) refer to as the fieldness of an experiment. For example, an experiment on voting may enhance ecological validity by being conducted in an actual polling place, using polling place equipment and registered voters.

Definition 7.8 (Ecological Validity): *Whether the methods, materials, and settings of the research are similar to a given target environment.*

However, this may or may not enhance external validity of the results because the target environment may not generalize. For example, the polling place equipment used in one jurisdiction may be different from that used in another jurisdiction. Thus, this may actually decrease the applicability of the results to different populations that use different types of equipment. Increasing ecological validity for *one* target population does not necessarily mean that the results generalize to *another* population and setting. External validity can only be established by generalizing beyond the target population and any target environment or setting. That said, increasing ecological validity and mundane realism of an experiment may help motivate subjects. We discuss this further in Chapter 10. We also return to Harrison and List's concerns about artificiality in experiments in Section 8.2.4.

7.3.2 Establishing External Validity

Suppose a researcher has been able to successfully identify and estimate a causal inference about a target population, using either experimental or nonexperimental data. Assume, for the moment, that the researcher is not engaging in theory testing and, thus, the construct validity of the initial analysis is not relevant. How can that researcher establish that the causal inference is externally valid? Or, more precisely, is it possible to establish the external validity of a causal inference that is not based on a theoretical construct without further empirical study? Without further empirical study, a researcher can only conjecture or hypothesize that his or her result has external validity based on similar studies or assumptions about the relationship between the population initially analyzed and the new population to be considered.

Is it different if the result validates a theoretical prediction and has construct validity? Although having construct validity helps us build a more general theory and provides evidence of a more general theory, we still cannot use theory to establish external validity. External validity can be conjectured or hypothesized based on similar studies or assumptions about population similarities about any study, experimental or nonexperimental, but the *proof* of external validity is always *empirical*. Debates about external validity in the absence of such empirical proof are debates about the similarity of a study to previous studies or population similarities, but there can never be a resolution through debate or discussion alone. Researchers would be better served by conducting more empirical studies than by debating external validity in the absence of such studies.

What sort of empirical analysis is involved in establishing external validity? A researcher simply replicates the empirical results on new populations or using new variations on the experiment in terms of settings, materials, and so on. With respect to establishing the external validity of results from theory evaluations, the researcher may also test new implications of the theory on the new populations as well as the old population. We discuss these processes later in this chapter.

Scientific Replication

Scientific replication is all about establishing external validity. It is when a researcher uses either a different sample or a different population to evaluate the same theoretical constructions with the same theoretical implications or uses the same or a different sample from either the same or a different population to evaluate different theoretical implications from these constructs. It is obviously less easily mandated by journals than statistical replication because it involves taking the same theoretical constructs and applying them to new populations or evaluating new theoretical implications or taking causal inferences based on fact searching and determining if they can be identified and estimated in a different data set. Often a researcher has used considerable effort to find, build, or create, as in an experiment, the data set for a study of a target population. Usually a researcher has sought all the data that he or she could find that was relevant and leaves establishing external validity through scientific replication to other researchers.

Definition 7.9 (Scientific Replication): *When a researcher uses a different sample, a different population to evaluate the same theoretical constructs with the same theoretical implications, or the same or a different sample from either the same or a different population to evaluate different theoretical implications from these constructs.*

One possible way to establish some external validity for one's own empirical results is through the use of *nonrandom holdout samples* as advocated by Keane and Wolpin (2007) and Wolpin (2007). A nonrandom holdout sample is one that differs significantly from the sample used for the estimation along a dimension over which the causal inference or theoretical prediction is expected to hold. If the empirical results from the original estimation are supported with the nonrandom holdout sample, which involves observations that are well outside the support of the original data, then the results will have more external validity along this dimension. As Keane and Wolpin remark, this procedure is often used in time-series analyses and has been used in the psychology and marketing literature. They note that such a procedure was used by McFadden (1977). McFadden estimated a random utility model of travel demand in the San Francisco Bay area before the introduction of the subway system and then compared his estimates to the actual usage after the subway was introduced. The observations after the subway was introduced composed the nonrandom holdout sample. Keane and Wolpin point out that experiments can provide an ideal opportunity for analyses with nonrandom holdout samples. One can imagine that treatments can be used as subsets of the population just as in the aforementioned cross-valuation procedure. Suppose a researcher conducts K treatments on different dimensions. Then the researcher can estimate the effects of the treatments on each of the possible groups of $K - 1$ subsets as separate target populations and then assess the predictive accuracy on the subset omitted on the dimension omitted. In this fashion, the researcher can gain some traction on the external validity of his or her results.

Definition 7.10 (Nonrandom Holdout Sample): *A nonrandom holdout sample is a sample that differs significantly from the sample used for the estimation along a dimension over which the causal inference or theoretical prediction is expected to hold.*

Although it is rare for a researcher to engage in scientific replication of his or her own research as described earlier, fortunately a lot of political science research does involve this sort of replication of the research of others. Gerber and Green's voter mobilization study was a scientific replication of the original study of Gosnell and the work of Rosenstone, as discussed previously.

Scientific replication through experimentation can occur when subjects from a different target population are used with the same experimental protocols to evaluate the same theoretical implications, or subjects from the same or different target population are used to evaluate different

theoretical implications sometimes with a change in experimental protocols (maintaining the same theoretical constructs). For example, Potters and van Winden (2000) replicated an experiment they had conducted previously with undergraduate students (Potters and van Winden, 1996), using lobbyists. One advantage of laboratory experiments is that usually statistical verification with different samples from the same target population can be reasonably conducted as long as researchers make publicly available detailed experimental protocols. Such explicit publicly available protocols are also required for effective scientific replications, particularly if the experimenter seeks to replicate with a sample from a new target population using the same experimental design. It is generally the norm of experimentalists in political science to provide access to these protocols for such replication. We believe this should be required of all political science experimentalists.

Stress Tests and External Validity

Recall that in Chapter 6 we referred to a type of experiment called a stress test as part of FTA. A stress test is also a way for an experimentalist to explore issues of external validity when evaluating a formal model. For example, suppose a researcher has tested a theory of legislative bargaining in the laboratory. The model is one of complete information. However, the researcher relaxes some of the information available to the subjects to determine if the behavior of the subjects will be affected. The researcher has no theoretical prediction about what will happen. If the theory's predictions hold despite this new wrinkle, then the researcher has learned that the results of the first experiment can generalize, under some circumstances, to a less than complete information environment. The experimental results are robust to this change if the theory's predictions hold. If the theory's predictions do not hold, then the experimental results are not robust.

Another example would be to conduct the same complete-information legislative bargaining theory experiment with different subject pools by conducting what is called lab-in-the-field versions of the experiment (discussed in Section 8.2.3) to determine how robust the results are to changes in who participates in the experiment. Again, if the theory's predictions hold, we say that the results are robust to this change, and vice versa. Or the experimentalist may vary the frame of the experiment – perhaps the original experiment used a neutral frame and subjects were told they were players in a game without any political context. The experimentalist could introduce a political context to the experiment by telling the subjects they are legislators and they are bargaining for ministerial positions and see if this frame difference affects the subjects' choices.

As noted in Chapter 6, the beauty of stress tests is that the experimenta can incorporate new features of the experimental environment on a pie meal basis and investigate each aspect of the change in an effort to test limits of the external robustness or validity of the results. Stress tests, th are important tools for experimentalists to test whether their results externally valid or robust and where in particular the robustness or valid may break down.

Analyses of Multiple Studies

Narrative and Systematic Reviews. The tendency of researchers in politi science is to look for new theoretical constructs or new theoretical imp cations from previously evaluated constructs that then become the fo of new empirical research. Alternatively, political scientists look for n target populations to evaluate existing theoretical constructs or establish causal relations. Much less often do political scientists conduct reviews research focusing on a particular research question. Yet, such reviews c be important in establishing the external validity of empirical results. the psychology and medical literature, these types of syntheses have beco commonplace to the extent that there is now a growing literature that repo on reviews of reviews.[8] Furthermore, many of the reviews in the psycholo and medical literature are quantitative in nature, using statistical metho to synthesize the results from a variety of studies, which are called me analysis, a term coined by Glass (1976). Researchers in the medical field a distinguish between a purely narrative review and a systematic review th includes both a narrative review and an analysis of the studies, either qu itative or quantitative. In this perspective a meta-analysis is a quantitat systematic review.

Definition 7.11 (Narrative Review): *Reviews of existing literature focusi on a particular research question.*

Definition 7.12 (Systematic Review): *A narrative review that includes eith a qualitative or quantitative synthesis of the reviewed studies' results.*

Definition 7.13 (Meta-analysis): *A quantitative systematic review using st tistical methods for which the researcher uses study results as the unit of obse vation or to construct the unit of observation.*

[8] For reviews of the literature on meta-analysis in other disciplines, see the special issue of t *International Journal of Epidemiology* in 2002, Bangert-Downs (1986), Delgado-Rodrigu (2006), Egger and Smith (1997), and Montori et al. (2003).

Political scientists sometimes use the term meta-analysis to refer to a literature review that is mainly narrative and qualitative. Political scientists also sometimes call a study that combines a couple of different empirical studies to address a single question, such as combining a laboratory experiment with a larger survey, a meta-analysis. Technically, neither are considered meta-analyses. In meta-analysis, usually the unit of observation is either the results of an overall study or results from distinctive parts of the study. Sometimes in meta-analysis researchers use statistical results from an overall study or distinctive parts to "approximate" data pooling (see Bangert-Downs, 1986). Other times, researchers actually pool all the data from multiple studies in cases where such data are available; such analyses are not usually considered meta-analyses but simply pooled analyses. In meta-analyses the researcher works with the reported information from the study which, of course, is secondary information, and this information serves as the basis of his or her statistical analysis. We expect that, as more political scientists begin to conduct systematic quantitative reviews as found in other disciplines, meta-analysis will have the same meaning in political science that it has in other disciplines, so we define a meta-analysis more narrowly.[9]

Definition 7.14 (Pooled Analysis): *A quantitative study that pools data from multiple studies to examine a particular research question.*

Issues in Meta-analyses. In a meta-analysis a researcher first has to decide on the criteria for including a study. Setting the criteria raises a lot of questions for the researcher. For example, suppose that the researcher is more suspect of the statistical or causal validity of some studies than others; should the researcher include all studies, but use statistics to control for these differences, or simply exclude studies with less valid results? As Bangert-Downs (1986) discussed, in psychology there has been much debate over whether low-quality studies should be included in meta-analysis – whether meta-analysis is simply "garbage in-garbage out" in such cases. Consider, for example, a meta-analysis that includes some experimental studies where

[9] A number of researchers have conducted systematic reviews that they call meta-analyses with case study data using case study methods. See, for example, Strandberg's (2008) study of the relationship between party Web sites and online electoral competition and Sager's (2006) study of policy coordination in European cities. Both of these studies use a method that has developed in case study research called qualitative comparative analysis (QCA). Because our focus in this book is on quantitative research taking an experimental approach, we do not include QCA approaches in our analysis.

causal validity is high with some nonexperimental studies where causal validity is not as high. Is it profitable to combine such studies for a meta-analysis? Alternatively, suppose that some of the data come from an experiment where random assignment has been utilized but another data set comes from an experiment without random assignment?

Studies also vary in the types of treatments and manipulations considered. Suppose that the treatment in a study is similar to the treatments given in other studies, but distinctive; to what extent can dissimilar studies be combined in an analysis that makes theoretical sense? One of the more seminal meta-analyses in psychology is Smith and Glass's (1977) study of the effects of psychotherapy. In this study the authors combined studies of a wide variety of psychotherapy from gestalt therapy to transactional analysis. What does such research tell us when so many different types of psychotherapy are combined? This is called the "apples-and-oranges" problem of meta-analysis. We could argue that doing so provides some overall measure of the effect of psychotherapy for policy makers who are choosing whether to support such therapies in general, but then what if one particular type of psychotherapy has been studied more often than it has actually been used, or has a bigger effect than others; does that skew the implications of the analysis?

After deciding on what types of studies to include, the researcher then faces additional statistical questions. What measures from the different studies should the research compare? Should the researcher compare significance and probabilities or sizes of effects? How does the researcher deal with publication biases? That is, suppose that studies showing no results or negative results are less likely to be published. How can the reviewer find information on such studies or, in the absence of such information, control for the possibility that they exist? Or, for example, suppose that the studies differ substantially in sample sizes, which has implications for comparisons across studies. How can a researcher control for these differences? Are there statistical techniques to estimate how robust the results of the reported studies are to unreported negative results? What happens if there is statistical dependence across different output measures?

Fortunately, the statistical methods used in meta-analysis are advanced enough in medicine and in psychology that researchers in political science who would like to conduct a meta-analysis can find a large literature on the methods that have been used to address these and many other methodological concerns. There are a number of textbooks on the subject (see, e.g., Hunter and Schmidt 1990). SCC also discussed meta-analysis at length in their Chapter 13. However, given the research interests of the other

disciplines, sometimes their answers are not appropriate for political science questions because many of the questions in medicine and psychology focus on particular isolated treatment effects of manipulations on individuals, whereas much of political science research examines effects at both individual and group levels and the interactions between the two. Furthermore, in the other disciplines, especially in medicine, it is likely that there are many studies that examine a common treatment and can be easily placed on a common metric for quantitative analyses, whereas doing so in political science may be more problematic.

Meta-analyses in Political Science. It is not surprising to us that meta-analyses are still rare in political science, mainly because it is difficult to think of a research question that has been the subject of the large number of studies needed for the statistical assumptions necessary for good meta-analysis. To our knowledge, meta-analyses have appeared at this writing only three times in the top three journals in political science, once in the *American Political Science Review* (Lau et al., 1999), once in the *Journal of Politics* (Lau et al., 2007, which is a replication of Lau et al., 1999), and once in the *American Journal of Political Science* (Doucouliagos and Ulubasoglu, 2008). Examples of meta-analyses are more numerous in specialized journals on public opinion and political psychology.

The meta-analyses of Lau et al. (1999, 2007) are instructive of how such synthesizing can lead to a deeper understanding of empirical relationships and provide insight into the complex choices facing researchers in meta-analyses. In these two studies the authors consider the empirical evidence on the effects of negative campaign advertising and find little support for the common perception in journalist circles that negative campaign advertising increases the probabilities that voters will choose the candidates who choose this strategy. As discussed earlier, the first step in the research approach used by Lau et al. (2007) was to decide on the criteria with which to include a study in their analysis. They chose to include studies that examined both actual and hypothetical political settings in which candidates or parties competed for support. Thus, they excluded studies of negative advertising in nonpolitical settings or in nonelectoral settings, but included studies for which the candidates and parties were hypothetical. If a researcher had reanalyzed previous data, they used the latest such study; however, they included studies by different researchers using different methods that used the same data set. Lau et al. also required that the study contain variation in the tone of the ads or campaigns. They focused on both studies that examined voter responses to the ads as intermediate effects as well as

their main interest on direct electoral effects and broader consequences on political variables such as turnout, voters' feelings of efficacy, trust, and political mood. The authors contend that these choices reflect their goal of answering the research question as to the effects of negative advertising in election campaigns. Yet, one may easily construct a meta-analysis that takes alternative focuses and uses different criteria. Ideally a researcher should consider how their criteria matter for the results provided. Lau et al. did consider the effects of using different studies from the same data set.

The second step in Lau et al.'s analysis was to do an extensive literature search to find all the relevant studies. Beyond simply surveying the literature, they contacted researchers working in the area, considered papers presented at conferences, and so on. This is a critical step in meta-analysis because it is important to avoid the "file drawer" problem of unpublished but important studies. The third step is to determine the measure for the quantitative analysis. Lau et al. focused on what is a standard technique in the psychology and medical literature, what is called Cohen's d or the *standardized mean difference statistic*, which is simply the difference in the means of the variable of interest in the treatment of interest versus the alternative treatment (or control group) divided by the pooled standard deviation of the two groups. Formally:

$$d_i = \frac{\overline{X_i^t} - \overline{X_i^c}}{s_i}, \tag{7.1}$$

where d_i is the standardized mean difference statistic for study i, $\overline{X_i^t}$ is the mean of the treatment group in the ith study, $\overline{X_i^c}$ is the mean of the control group in the ith study, and s_i is the pooled standard deviation of the two groups.

In experiments, the d statistic is relatively easy to calculate if a researcher has knowledge of the sample sizes and the standard deviations of the two groups being compared. However, if some of the studies contain nonexperimental data and are multivariate analyses, the researcher may not be able to easily calculate these measures. Lau et al. used an approximation for d in such cases that is derived from the t statistic, suggested by Stanley and Jarrell (1989), which is called by Rosenthal and Rubin (2003) the $d_{\text{equivalent}}$. Formally:

$$d_{\text{equivalent}} = \frac{2t}{\sqrt{df}}, \tag{7.2}$$

where t is the t statistic from the multivariate regression for the independent variable of interest and df is the degrees of freedom associated with the t

test. In their appendix, Lau et al. (1999) describe this measure in detail. This measure, of course, assumes that the independent variable associated with the t statistic is an accurate measure of the causal effect that the meta-analysis is studying. The important implicit assumptions implied by the use of this measure are explored more expansively in Chapter 5, when we discuss how causal inferences can be estimated from nonexperimental data.

After calculating the values of d, Lau et al. also had to deal with the fact that the different data sets combine different sample sizes. A number of methods exist in the literature to adjust for sample sizes (see, e.g., Hedges and Olkin, 1985). Lau et al. (1999, 2007) used a method recommended by Hunter and Schmidt (1990) to weight for sample size differences. These weights are described in the appendix to Lau et al. (1999). The authors also adjusted their measure for reliability of the variables as recommended by Hunter and Schmidt. For those outcomes for which studies report reliability, they used that measure; for studies that did not report reliability measures, they used the mean reliability for other findings within the same dependent-variable category. Finally, the authors adjusted the data for variability in the strength of the negative advertisement "treatments."

Shadish and Haddock (1994) noted that, in cases where all the studies considered use the same outcome measure, it might make sense to use the difference between raw means as the common metric. In other cases, the researcher may not be examining mean differences at all. For example, Oosterbeek et al. (2004) conducted a meta-analysis of choices of subjects in an experimental bargaining game. The analysis was a study of the determinants of the size of proposals made in the bargaining and the probability that proposals are rejected. There was no treatment or baseline in these experiments in the traditional sense because the question of interest is the extent that subjects deviate from theoretical point predictions rather than a comparative static prediction. Because the size of the bargaining pies varied as well as the relative values, Oosterbeek et al. controlled for such differences. We discuss this study more expansively in the next chapter because the study considers the effects of different subject pools in laboratory experiments.

An alternative to the d measure is the correlation coefficient as the effect size. For example, Doucouliagos and Ulubasoglu (2008) used partial correlations as their effect size measures weighted for sample size (see discussion of this measure by Ones et al. 1993; Rosenthal and Rubin, 1978). It makes sense where the studies reviewed examine the same correlational relationship among variables. Greene (2000, p. 234) provides details on how to calculate partial correlations from regression outputs of studies.

Dougcouliagos and Ulubascoglu controlled for variations across the studies they examined in the empirical analysis of the data.

The d measure assumes that the study outcome is measured continuously. If the study outcome is binary, then d can yield problematic effect size estimates (see Fleiss, 1994; Haddock et al., 1998). In this case the effect size can be measured by the odds ratio. Formally:

$$o_i = \frac{AD}{BC}, \tag{7.3}$$

where o_i is the odds ratio for study i, A is the frequency with which the treatment occurs and there is no effect on the outcome, B is the frequency with which the treatment occurs and there is an effect on the outcome, C is the frequency with which the treatment is absent and there is no effect on the outcome, and D is the frequency with which the treatment is absent and there is an effect on the outcome.

Clearly, all of the decisions that researchers like Lau et al. make in conducting a meta-analysis affect the validity of the results. SCC and Hunter and Schmidt discuss these issues in detail.

7.3.3 Is External Validity Possible Without Satisfying Internal Validity?

Many political scientists quickly concede that experimental research has high internal validity compared with research with observational data and they dismiss experimental research (especially laboratory experiments) as being low on external validity compared with research with observational data. Both opinions tend to understate and ignore the multitude of issues involved in establishing the internal validity of both observational and experimental research, which we have discussed in this chapter. In particular, in our view, external validity can only be established for results that have been demonstrated to be internally valid in the senses we have mentioned – statistical conclusion, causal validity, and, if the empirical study involves theory testing, construct validity. If a result is not statistically significant, cannot be established to be causal in the population originally investigated, or is estimated from an empirical study that has little relevance to the theory being evaluated, then how can it possibly be considered externally valid or robust as a causal relationship? It makes no sense to say that some empirical research is low on internal validity but high on external validity.