FIELD EXPERIMENTS

Design, Analysis, and Interpretation

Alan S. Gerber YALE UNIVERSITY

Donald P. Green COLUMBIA UNIVERSITY

W. W. NORTON & COMPANY NEW YORK • LONDON

W. W. Norton & Company has been independent since its founding in 1923, when William Warder Norton and Mary D. Herter Norton first published lectures delivered at the People's Institute, the adult education division of New York City's Cooper Union. The firm soon expanded its program beyond the Institute, publishing books by celebrated academics from America and abroad. By midcentury, the two major pillars of Norton's publishing program—trade books and college texts—were firmly established. In the 1950s, the Norton family transferred control of the company to its employees, and today—with a staff of four hundred and a comparable number of trade, college, and professional titles published each year—W. W. Norton & Company stands as the largest and oldest publishing house owned wholly by its employees.

Editor: Ann Shin
Associate Editor: Jake Schindel
Project Editor: Jack Borrebach
Marketing Manager, political science: Sasha Levitt
Production Manager: Eric Pier-Hocking
Text Design: Joan Greenfield / Gooddesign Resource
Design Director: Hope Miller Goodell
Composition by Jouve International—Brattleboro, VT
Manufacturing by the Maple Press—York, PA

Printed in the United States of America.
First edition.

Library of Congress Cataloging-in-Publication Data

Gerber, Alan S.
Field experiments : design, analysis, and interpretation / Alan S. Gerber, Donald P. Green. — 1st ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-393-97995-4 (pbk.)
1. Political science—Research—Methodology. 2. Social science—Research—Methodology.
3. Political science—Study and teaching (Higher) 4. Social science—Study and teaching (Higher)
I. Green, Donald P., 1961-II. Title.
JA86.G36 2012
001.4'34—dc23
2011052337

W. W. Norton & Company, Inc., 500 Fifth Avenue, New York, NY 10110-0017
wwnorton.com
W. W. Norton & Company Ltd., Castle House, 75/76 Wells Street, London W1T 3QT

1 2 3 4 5 6 7 8 9 0

CHAPTER 1

Introduction

Daily life continually presents us with questions of cause and effect. Will eating more vegetables make me healthier? If I drive a bit faster than the law allows, will the police pull me over for a speeding ticket? Will dragging my reluctant children to museums make them one day more interested in art and history? Even actions as banal as scheduling a dental exam or choosing an efficient path to work draw on cause-and-effect reasoning.

Organizations, too, grapple with causal puzzles. Charities try to figure out which fundraising appeals work best. Marketing agencies look for ways to boost sales. Churches strive to attract congregants on Sundays. Political parties maneuver to win elections. Interest groups attempt to influence legislation. Whether their aim is to boost donations, sales, attendance, or political influence, organizations make decisions based (at least in part) on their understanding of cause and effect. In some cases, the survival of an organization depends on the skill with which it addresses the causal questions that it confronts.

Of special interest to academic researchers are the causal questions that confront governments and policy makers. What are the economic and social effects of raising the minimum wage? Would allowing parents to pay for private school using publicly funded vouchers make the educational system more effective and cost-efficient? Would legal limits on how much candidates can spend when running for office affect the competitiveness of elections? In the interest of preventing bloodshed, should international peacekeeping troops be deployed with or without heavy weapons? Would mandating harsher punishments for violent offenders deter crime? A list of policy-relevant causal questions would itself fill a book.

An even larger tome would be needed to catalog the many theoretical questions that are inspired by causal claims. For example, when asked to contribute to a collective cause, such as cutting down on carbon emissions in order to prevent global climate change, to what extent are people responsive to appeals based on social norms or ideology? Prominent scholars have argued that collective action will founder

unless individuals are given some sort of reward for their participation; according to this argument, simply telling people that they ought to contribute to a collective cause will not work.[1] If this underlying causal claim is true, the consequences for policymaking are profound: tax credits may work, but declaring a national Climate Change Awareness Day will not.

Whether because of their practical, policy, or theoretical significance—or simply because they transport us to a different time and place—causal claims spark the imagination. How does the pilgrimage to Mecca affect the religious, social, and political attitudes of Muslims?[2] Do high school dropout rates in low-income areas improve when children are given monetary rewards for academic performance?[3] Are Mexican police more likely to demand bribes from upper- or lower-class drivers who are pulled aside for traffic infractions?[4] Does your race affect whether employers call you for a job interview?[5] In the context of a civil war, do civilians become more supportive of the government when local economic conditions improve?[6] Does artillery bombardment directed against villages suspected of harboring insurgent guerrillas increase or decrease the likelihood of subsequent insurgent attacks from those villages?[7]

In short, the world is brimming over with causal questions. How might one go about answering them in a convincing manner? What methods for answering causal questions should be viewed with skepticism?

1.1 Drawing Inferences from Intuitions, Anecdotes, and Correlations

One common way of addressing causal questions is to draw on intuition and anecdotes. In the aforementioned case of artillery directed at insurgent villages, a scholar might reason that firing on these villages could galvanize support for the rebels, leading to more insurgent attacks in the future. Bombardment might also prompt the rebels to demonstrate to villagers their determination to fight on by escalating their insurgent activities. In support of this hypothesis, one might point out that the anti-Nazi insurgency in Soviet Russia in 1941 became more determined after occupation forces stepped up their military suppression. One problem with building causal arguments

1 Olson 1965.
2 Clingingsmith, Khwaja, and Kremer 2009.
3 Angrist and Lavy 2009; see also Fryer 2010.
4 Fried, Lagunes, and Venkataramani 2010.
5 Bertrand and Mullainathan 2004.
6 Beath, Christia, and Enikolopov 2011.
7 Lyall 2009.

around intuitions and anecdotes, however, is that such arguments can often be adduced for both sides of a causal claim. In the case of firing on insurgents, another researcher could argue that insurgents depend on the goodwill of villagers; once a village is fired upon, villagers have a greater incentive to expel the rebels in order to prevent future attacks. Supplies dry up, and informants disclose rebel hideouts to government forces. This researcher could defend the argument by describing the government suppression of the Sanusi uprising in Libya, which seemed to deal a lasting blow to these rebels' ability to carry out insurgent attacks.[8] Debates based on intuition and anecdotes frequently result in stalemate.

A critique of anecdote and intuition can be taken a step further. The method is susceptible to error even when intuition and anecdotes seem to favor just one side of an argument. The history of medicine, which is instructive because it tends to provide clearer answers to causal questions than research in social science, is replete with examples of well-reasoned hypotheses that later proved to be false when tested experimentally. Consider the case of aortic arrhythmia (irregular heartbeat), which is often associated with heart attacks. A well-regarded theory held that arrhythmia was a precursor to heart attack. Several drugs were developed to suppress arrhythmia, and early clinical reports seemed to suggest the benefits of restoring a regular heartbeat. The Cardiac Arrhythmia Suppression Trial, a large randomized experiment, was launched in the hope of finding which of three suppression drugs worked best, only to discover that two of the three drugs produced a significant *increase* in death and heart attacks, while the third had negative but seemingly less fatal consequences.[9] The broader point is that well-regarded theories are fallible. This concern is particularly acute in the social sciences, where intuitions are rarely uncontroversial, and controversial intuitions are rarely backed up by conclusive evidence.

Another common research strategy is to assemble statistical evidence showing that an outcome becomes more likely when a certain cause is present. Researchers sometimes go to great lengths to assemble large datasets that allow them to track the correlation between putative causes and effects. These data might be used to learn about the following statistical relationship: to what extent do villages that come under attack by government forces tend to have more or less subsequent insurgent activity? Sometimes these analyses turn up robust correlations between interventions and outcomes. The problem is that correlations can be a misleading guide to causation. Suppose, for example, that the correlation between government bombardment and subsequent insurgent activity were found to be strongly positive: the more shelling, the more subsequent insurgent activity. If interpreted causally, this correlation would indicate that shelling prompted insurgents to step up their attacks. Other

8 See Lyall 2009 for a discussion of these debates and historical episodes.
9 Cardiac Arrhythmia Suppression Trial II Investigators 1992.

interpretations, however, are possible. It could be that government forces received intelligence about an escalation of insurgent activity in certain villages and directed their artillery there. Shelling, in other words, could be a marker for an uptick in insurgent activity. Under this scenario, we would observe a positive correlation between shelling and subsequent insurgent attacks even if shelling per se had no effect.

The basic problem with using correlations as a guide to causality is that correlations may arise for reasons that have nothing to do with the causal process under investigation. Do SAT preparation courses improve SAT scores? Suppose there were a strong positive correlation here: people who took a prep class on average got higher SAT scores than those who did not take the prep class. Does this correlation reflect the course-induced improvement in scores, or rather the fact that students with the money and motivation to take a prep course tend to score higher than their less affluent or less motivated counterparts? If the latter were true, we might see a strong association even if the prep course had no effect on scores. A common error is to reason that where there's smoke, there's fire: correlations at least hint at the existence of a causal relationship, right? Not necessarily. Basketball players tend to be taller than other people, but you cannot grow taller by joining the basketball team.

The distinction between correlation and causation seems so fundamental that one might wonder why social scientists rely on correlations when making causal arguments. The answer is that the dominant methodological practice is to transform raw correlations into more refined correlations. After noticing a correlation that might have a causal interpretation, researchers attempt to make this causal interpretation more convincing by limiting the comparison to observations that have similar background attributes. For example, a researcher seeking to isolate the effects of the SAT preparatory course might restrict attention to people with the same gender, age, race, grade point average, and socioeconomic status. The problem is that this method remains vulnerable to *unobserved* factors that predict SAT scores and are correlated with taking a prep course. By restricting attention to people with the same socio-demographic characteristics, a researcher makes the people who took the course comparable to those who did not in terms of *observed* attributes, but these groups may nevertheless differ in ways that are unobserved. In some cases, a researcher may fail to consider some of the factors that contribute to SAT scores. In other cases, a researcher may think of relevant factors but fail to measure them adequately. For example, people who take the prep course may, on average, be more motivated to do well on the test. If we fail to measure motivation (or fail to measure it accurately), it will be one of the unmeasured attributes that might cause us to draw mistaken inferences. These unmeasured attributes are sometimes called *confounders* or *lurking variables* or *unobserved heterogeneity*. When interpreting correlations, researchers must always be alert to the distorting influence of unmeasured attributes. The fact that someone chooses to take the prep course may reveal something about how they are likely to perform on the test. Even if the course truly has no effect, people with the same age, gender, and affluence may seem to do better when they take the course.

Whether the problem of unobserved confounders is severe or innocuous will depend on the causal question at hand and the manner in which background attributes are measured. Consider the so-called "broken windows" theory, which suggests that crime increases when blighted areas appear to be abandoned by property owners and unsupervised by police.[10] The causal question is whether one could reduce crime in such areas by picking up trash, removing graffiti, and repairing broken windows. A weak study might compare crime rates on streets with varying levels of property disrepair. A more convincing study might compare crime rates on streets that currently experience different levels of blight but in the past had similar rates of disrepair and crime. But even the latter study may still be unconvincing because unmeasured factors, such as the closing of a large local business, may have caused some streets to deteriorate physically and coincided with an upsurge in crime.[11]

Determined to conquer the problem of unobserved confounders, one could set out to measure each and every one of the unmeasured factors. The intrepid researcher who embarks on this daunting task confronts a fundamental problem: no one can be sure what the set of unmeasured factors comprises. The list of all potential confounders is essentially a bottomless pit, and the search has no well-defined stopping rule. In the social sciences, research literatures routinely become mired in disputes about unobserved confounders and what to do about them.

1.2 Experiments as a Solution to the Problem of Unobserved Confounders

The challenge for those who seek to answer causal questions in a convincing fashion is to come up with a research strategy that does not require them to identify, let alone measure, all potential confounders. Gradually, over the course of centuries, researchers developed procedures designed to sever the statistical relationship between the treatment and all variables that predict outcomes. The earliest experiments, such as Lind's study of scurvy in the 1750s and Watson's study of smallpox in the 1760s, introduced the method of systematically tracking the effects of a researcher-induced intervention by comparing outcomes in the treatment group to outcomes in one or more control groups.[12] One important limitation of these early studies is that they assumed that their subjects were identical in terms of their medical trajectories. What if this assumption

10 Wilson and Kelling 1982.

11 See Keizer, Lindenberg, and Steg 2008, but note that this study does not employ random assignment. For a randomized field experiment see Mazerolle, Price, and Roehl 2000.

12 Hughes 1975; Boylston 2008.

were false, and treatments tended to be administered to patients with the best chances of recovery? Concerned that the apparent effects of their intervention might be attributable to extraneous factors, researchers placed increasing emphasis on the procedure by which treatments were assigned to subjects. Many pathbreaking studies of the nineteenth century assigned subjects alternately to treatment and control in an effort to make the experimental groups comparable. In 1809, a Scottish medical student described research conducted in Portugal in which army surgeons treated 366 sick soldiers alternately with bloodletting and other palliatives.[13] In the 1880s, Louis Pasteur tested his anthrax vaccine on animals by alternately exposing treatment and control groups to the bacteria. In 1898, Johannes Fibiger assigned an experimental treatment to diphtheria patients admitted to a hospital in Copenhagen on alternate days.[14] Alternating designs were common in early agricultural studies and investigations of clairvoyance, although researchers gradually came to recognize potential pitfalls of alternation.[15] One problem with alternating designs is that they cannot definitively rule out confounding factors, such as sicker diphtheria patients coming to the hospital on certain days of the week. The first to recognize the full significance of this point was the agricultural statistician R. A. Fisher, who in the mid-1920s argued vigorously for the advantages of assigning observations at random to treatment and control conditions.[16]

This insight represents a watershed moment in the history of science. Recognizing that no planned design, no matter how elaborate, could fend off every possible systematic difference between the treatment and control groups, Fisher laid out a general procedure for eliminating systematic differences between treatment and control groups: random assignment. When we speak of experiments in this volume, we refer to studies in which some kind of random procedure, such as a coin flip, determines whether a subject receives a treatment.

One remarkable aspect of the history of randomized experimentation is that the idea of random assignment occurred to several ingenious people centuries before it was introduced into modern scientific practice. For example, the notion that one could use random assignment to form comparable experimental groups seems to have been apparent to the Flemish physician Jan Baptist Van Helmont, whose 1648 manuscript "Origin of Medicine" challenged the proponents of bloodletting to perform the following randomized experiment:

> Let us take out of the hospitals . . . 200 or 500 poor people, that have fevers, pleurisies. Let us divide them into halves, let us cast lots, that one halfe of them may fall to my share, and the other to yours; I will cure them without bloodletting and sensible evacuation; but you do, as ye know . . . We shall see how many funerals both of us shall have.[17]

13 Chalmers 2001.

14 Hróbjartsson, Gøtzsche, and Gluud 1998.

15 Merrill 2010. For further reading on the history of experimentation, see Cochran 1976; Forsetlund, Chalmers, and Bjørndal 2007; Hacking 1990; and Salsburg 2001. See Greenberg and Shroder 2004 on social experiments and Green and Gerber 2003 on the history of experiments in political science.

16 Box 1980, p. 3.

Unfortunately for those whose physicians prescribed bloodletting in the centuries following Van Helmont, he never conducted his proposed experiment. One can find similar references to hypothetical experiments dating back to medieval times, but no indication that any were actually put into practice. Until the advent of modern statistical theory in the early twentieth century, the properties of random assignment were not fully appreciated, nor were they discussed in a systematic manner that would have allowed one generation to recommend the idea to the next.

Even after Fisher's ideas became widely known in the wake of his 1935 book *The Design of Experiments*, randomized designs met resistance from medical researchers until the 1950s, and randomized experiments did not catch on in the social sciences until the 1960s.[18] In the class of brilliant twentieth-century discoveries, the idea of randomization contrasts sharply with the idea of relativity, which lay completely hidden until uncovered by genius. Randomization was more akin to crude oil, something that periodically bubbled to the surface but remained untapped for centuries until its extraordinary practical value came to be appreciated.

1.3 Experiments as Fair Tests

In the contentious world of causal claims, randomized experimentation represents an evenhanded method for assessing what works. The procedure of assigning treatments at random ensures that there is no systematic tendency for either the treatment or control group to have an advantage. If subjects were assigned to treatment and control groups and no treatment were actually administered, there would be no reason to expect that one group would outperform the other. In other words, random

17 Chalmers 2001, p. 1157.

18 The advent of randomized experimentation in social and medical research took roughly a quarter century. Shortly after laying the statistical foundations for random assignment and the analysis of experimental data, Fisher collaborated on the first randomized agricultural experiment (Eden and Fisher 1927). Within a few years, Amberson, McMahon, and Pinner (1931) performed what appears to be the first randomized medical experiment, in which tuberculosis patients were assigned to clinical trials based on a coin flip. The large-scale studies of tuberculosis conducted during the 1940s brought randomized clinical trials to the forefront of medicine. Shortly afterward, the primacy of this methodology in medicine was cemented by a series of essays by Hill (1951, 1952) and subsequent acclaim of the polio vaccine trials of the 1950s (Tanur 1989). Randomized clinical trials gradually came to be heralded as the gold standard by which medical claims were to be judged. By 1952, books such as Kempthorne's *Design and Analysis of Experiments* (pp. 125–126) declared that "only when the treatments in the experiment are applied by the experimenter using the full randomization procedure is the chain of inductive inference sound."

assignment implies that the observed *and unobserved* factors that affect outcomes are equally likely to be present in the treatment and control groups. Any given experiment may overestimate or underestimate the effect of the treatment, but if the experiment were conducted repeatedly under similar conditions, the average experimental result would accurately reflect the true treatment effect. In Chapter 2, we will spell out this feature of randomized experiments in greater detail when we discuss the concept of unbiased estimation.

Experiments are fair in another sense: they involve transparent, reproducible procedures. The steps used to conduct a randomized experiment may be carried out by any research group. A random procedure such as a coin flip may be used to allocate observations to treatment or control, and observers can monitor the random assignment process to make sure that it is followed faithfully. Because the allocation process precedes the measurement of outcomes, it is also possible to spell out beforehand the way in which the data will be analyzed. By automating the process of data analysis, one limits the role of discretion that could compromise the fairness of a test.

Random allocation is the dividing line that separates experimental from nonexperimental research in the social sciences. When working with nonexperimental data, one cannot be sure whether the treatment and control groups are comparable because no one knows precisely why some subjects and not others came to receive the treatment. A researcher may be prepared to assume that the two groups are comparable, but assumptions that seem plausible to one researcher may strike another as far-fetched.

This is not to say that experiments are free from problems. Indeed, this book would be rather brief were it not for the many complications that may arise in the course of conducting, analyzing, and interpreting experiments. Entire chapters are devoted to problems of noncompliance (subjects who receive a treatment other than the one to which they were randomly assigned), attrition (observations for which outcome measurements are unavailable), and interference between units (observations influenced by the experimental conditions to which other observations are assigned). The threat of bias remains a constant concern even when conducting experiments, which is why it is so important to design and analyze them with an eye toward maintaining symmetry between treatment and control groups and, more generally, to embed the experimental enterprise in institutions that facilitate proper reporting and accumulation of experimental results.

1.4 Field Experiments

Experiments are used for a wide array of different purposes. Sometimes the aim of an experiment is to assess a theoretical claim by testing an implied causal relationship. Game theorists, for example, use laboratory experiments to show how the introduction

BOX 1.1

Experiments in the Natural Sciences

Readers with a background in the natural sciences may find it surprising that random assignment is an integral part of the definition of a social science experiment. Why is random assignment often unnecessary in experiments in, for example, physics? Part of the answer is that the "subjects" in these experiments—e.g., electrons—are more or less interchangeable, and so the method used to assign subjects to treatment is inconsequential. Another part of the answer is that lab conditions neutralize all forces other than the treatment.

In the life sciences, subjects are often different from one another, and eliminating unmeasured disturbances can be difficult even under carefully controlled conditions. An instructive example may be found in a study by Crabbe, Wahlsten, and Dudek (1999), who performed a series of experiments on mouse behavior in three different science labs. As Lehrer (2010) explains:

> Before [Crabbe] conducted the experiments, he tried to standardize every variable he could think of. The same strains of mice were used in each lab, shipped on the same day from the same supplier. The animals were raised in the same kind of enclosure, with the same brand of sawdust bedding. They had been exposed to the same amount of incandescent light, were living with the same number of littermates, and were fed the exact same type of chow pellets. When the mice were handled, it was with the same kind of surgical glove, and when they were tested it was on the same equipment, at the same time in the morning.

Nevertheless, experimental interventions produced markedly different results across mice and research sites.

of uncertainty or the opportunity to exchange information prior to negotiating affects the bargains that participants strike with one another.[19] Such experiments are often couched in very abstract terms, with rules that stylize the features of an auction, legislative session, or international dispute. The participants are typically ordinary people (often members of the university community), not traders, legislators, or diplomats, and the laboratory environment makes them keenly aware that they are participating in a research study.

At the other end of the spectrum are experiments that strive to be as realistic and unobtrusive as possible in an effort to test more context-specific hypotheses.

19 See Davis and Holt 1993; Kagel and Roth 1995; Guala 2005.

Quite often this type of research is inspired by a mixture of theoretical and practical concerns. For example, to what extent and under what conditions does preschool improve subsequent educational outcomes? Experiments that address this question shed light on theories about childhood development while at the same time informing policy debates about whether and how to allocate resources to early childhood education in specific communities.

The push for realism and unobtrusiveness stems from the concern that unless one conducts experiments in a naturalistic setting and manner, some aspect of the experimental design may generate results that are idiosyncratic or misleading. If subjects know that they are being studied or if they sense that the treatment they received is supposed to elicit a certain kind of response, they may express the opinions or report the behavior they believe the experimenter wants to hear. A treatment may seem effective until a more unobtrusive experiment proves otherwise.[20] Conducting research in naturalistic settings may be viewed as a hedge against unforeseen threats to inference that arise when drawing generalizations from results obtained in laboratory settings. Just as experiments are designed to test causal claims with minimal reliance on assumptions, experiments conducted in real-world settings are designed to make generalizations less dependent on assumptions.

Randomized studies that are conducted in real-world settings are often called *field experiments*, a term that calls to mind early agricultural experiments that were literally conducted in fields. The problem with the term is that the word *field* refers to the setting, but the setting is just one aspect of an experiment. One should invoke not one but several criteria: whether the treatment used in the study resembles the intervention of interest in the world, whether the participants resemble the actors who ordinarily encounter these interventions, whether the context within which subjects

20 Whether this concern is justified is an empirical question, and the answer may well depend on the setting, context, and subjects. Unfortunately, the research literature on this topic remains underdeveloped. Few studies have attempted to estimate treatment effects in both lab and field contexts. Gneezy, Haruvy, and Yafe (2004), for example, use field and lab studies to test the hypothesis that the quantity of food consumed depends on whether each diner pays for his or her own food or whether they all split the bill. When this experiment is conducted in an actual cafeteria, splitting the bill leads to significantly more food consumption; when the equivalent game is played in abstract form (with monetary payoffs) in a nearby lab, the average effect is weak and not statistically distinguishable from zero. Jerit, Barabus, and Clifford (2012) compare the effects of exposure to a local newspaper on political knowledge and opinions. In the field, free Sunday newspapers were randomly distributed to households over the course of one month; in the lab, subjects from the same population were invited to a university setting, where they were presented with the four most prominent political news stories airing during the same month. For the 17 outcome measures, estimated treatment effects in the lab and field are found to be weakly correlated (Table 2). See also Rondeau and List (2008), who compare the effectiveness of different fundraising appeals on behalf of the Sierra Club directed at 3,000 past donors, as measured by actual donations. The fundraising appeals, which involve various combinations of matching funds, thresholds, and money-back guarantees, are then presented in abstract form in a lab setting with monetary payoffs. The correspondence between lab and field results was relatively weak, with average contributions in the lab predicting about 5% of the variance in average contributions in the field across the four conditions.

receive the treatment resembles the context of interest, and whether the outcome measures resemble the actual outcomes of theoretical or practical interest.

For example, suppose one were interested in the extent to which financial contributions to incumbent legislators' reelection campaigns buy donors access to the legislators, a topic of great interest to those concerned that the access accorded to wealthy donors undermines democratic representation. The hypothesis is that the more a donor contributes, the more likely the legislator is to grant a meeting to discuss the donor's policy prescriptions. One possible design is to recruit students to play the part of legislative schedulers and present them with a list of requests for meetings from an assortment of constituents and donors in order to test whether people described as potential donors receive priority. Another design involves the same exercise, but this time the subjects are actual legislative schedulers.[21] The latter design would seem to provide more convincing evidence about the relationship between donations and access in actual legislative settings, but the degree of experimental realism remains ambiguous. The treatments in this case are realistic in the sense that they resemble what an actual scheduler might confront, but the subjects are aware that they are participating in a simulation exercise. Under scrutiny by researchers, legislative schedulers might try to appear indifferent to fundraising considerations; in an actual legislative setting where principals provide feedback to schedulers, donors might receive special consideration. More realistic, then, would be an experiment in which one or more donors contribute randomly assigned sums of money to various legislators and request meetings to discuss a policy or administrative concern. In this design, the subjects are actual schedulers, the treatment is a campaign donation, the treatment and request for a meeting are authentic, and the outcome is whether a real request is granted in a timely fashion.

Because the degree of "fieldness" may be gauged along four different dimensions (authenticity of treatments, participants, contexts, and outcome measures), a proper classification scheme would involve at least sixteen categories, a taxonomy that far exceeds anyone's interest or patience. Suffice it to say that field experiments take many forms. Some experiments seem naturalistic on all dimensions. Sherman et al. worked with the Kansas City police department in order to test the effectiveness of police raids on locations where drug dealing was suspected.[22] The treatments were raids by teams of uniformed police directed at 104 randomly chosen sites among the 207 locations for which warrants had been issued. Outcomes were crime rates in nearby areas. Karlan and List collaborated with a charity in order to test the effectiveness of alternative fundraising appeals.[23] The treatments were fundraising letters; the experiment was unobtrusive in the sense that recipients of the fundraising appeals were

21 See Chin, Bond, and Geva 2000.
22 Sherman et al. 1995.
23 Karlan and List 2007.

unaware that an experiment was being conducted; and the outcomes were financial donations. Bergan teamed up with a grassroots lobbying organization in order to test whether constituents' e-mail to state representatives influences roll call voting.[24] The lobbying organization allowed Bergan to extract a random control group from its list of targeted legislators; otherwise, its lobbying campaign was conducted in the usual way, and outcomes were assessed based on the legislators' floor votes.

Many field experiments are less naturalistic, and generalizations drawn from them are more dependent on assumptions. Sometimes the interventions deployed in the field are designed by researchers rather than practitioners. Eldersveld, for example, fashioned his own get-out-the-vote campaigns in order to test whether mobilization activities cause registered voters to cast ballots.[25] Much may be learned when researchers craft their own treatments—indeed, the development of theoretically inspired interventions is an important way in which researchers may contribute to theoretical and policy debates. However, if the aim of an experiment is to gauge the effectiveness of typical candidate- or party-led voter mobilization campaigns, researcher-led campaigns may be unrepresentative in terms of the messages used or the manner in which they are communicated. Suppose that the researcher's intervention were to prove ineffective. This finding alone would not establish that a typical campaign's interventions are ineffective, although this interpretation could be bolstered by a series of follow-up experiments that test different types of campaign communication.[26] Sometimes treatments are administered and outcomes are measured in a way that notifies participants that they are being studied, as in Paluck's experimental investigation of intergroup prejudice in Rwanda.[27] Her study enlisted groups of Rwandan villagers to listen to recordings of radio programs on a monthly basis for a period of one year, at which point outcomes were measured using surveys and role-playing exercises. Finally, experimental studies with relatively little field content are those in which actual interventions are delivered in artificial settings to subjects who are aware that they are part of a study. Examples of this type of research may be found in the domain of commercial advertising, where subjects are shown different types of ads either in the context of an Internet survey or in a lab located in a shopping center.[28]

Whether a given study is regarded as a field experiment is partly a matter of perspective. Ordinarily, experiments that take place on college campuses are considered lab studies, but some experiments on cheating involve realistic opportunities for students to copy answers or misreport their own performance on self-graded tests.[29] An experimental study that examines the deterrent effect of exam proctoring would amount to a field experiment if one's aim were to understand the conditions under which students cheat in school. This example serves as a reminder that what constitutes a field experiment depends on how "the field" is defined.

24 Bergan 2009.

25 Eldersveld 1956.

26 For example, in an effort to test whether voter mobilization phone calls conducted by call centers are typically ineffective, Panagopoulos (2009) compares partisan and nonpartisan scripts, Nickerson (2007) assesses whether effectiveness varies depending on the quality of the calling center, and other scholars have conducted studies in various electoral environments. See Green and Gerber 2008 for a review of this literature.

27 Paluck 2009.

28 See, for example, Clinton and Lapinski 2004; Kohn, Smart, and Ogborne 1984.

1.5 Advantages and Disadvantages of Experimenting in Real-World Settings

Many field experiments take the form of "program evaluations" designed to gauge the extent to which resources are deployed effectively. For example, in order to test whether a political candidate's TV advertising campaign increases her popularity, a field experiment might randomize the geographic areas in which the ads are deployed and measure differences in voter support between treatment and control regions. From the standpoint of program evaluation, this type of experiment is arguably superior to a laboratory study in which voters are randomly shown the candidate's ads and later asked their views about the candidate. The field experiment tests the effects of deploying the ads and allows for the possibility that some voters in targeted areas will miss the ad, watch it inattentively, or forget its message amid life's other distractions. Interpretation of the lab experiment's results is complicated by the fact that subjects in lab settings may respond differently to the ads than the average voter outside the lab. In this application, preliminary lab research might be useful insofar as it suggests which messages are most likely to work in field settings, but only a field experiment allows the researcher to reliably gauge the extent to which an actual ad campaign changed votes and to express this outcome in relation to the resources spent on the campaign.

As we move from program evaluation to tests of theoretical propositions, the relative merits of field and lab settings become less clear-cut. A practical advantage of delivering treatments under controlled laboratory conditions is that one can more easily administer multiple variations of a treatment to test fine-grained theoretical propositions. Field interventions are often more cumbersome: in the case of political advertisements, it may be logistically challenging or politically risky to air multiple advertisements in different media markets. On the other hand, field experiments are sometimes able to achieve a high level of theoretical nuance when a wide array of treatments can be distributed across a large pool of subjects. Field experiments that deploy multiple versions of a treatment are common, for example, in research

29 Canning 1956; Nowell and Laufer 1997.

on discrimination, where researchers vary ethnicity, social class, and a host of other characteristics to better understand the conditions under which discrimination occurs.[30]

Even when limited to a single, relatively blunt intervention, a researcher may still have reason to conduct experiments in the field. Advertising research in field settings is often unobtrusive in the sense that subjects are not viewing the ad at the behest of a researcher, and outcomes are measured in a way that does not alert subjects to the fact that they are being studied.[31] Whereas outcomes in lab settings are often attitudes and behaviors that can be measured in the space of one sitting,[32] field studies tend to monitor behaviors over extended periods of time. The importance of ongoing outcome measurement is illustrated by experiments that find strong instantaneous effects of political advertising that decay rapidly over time.[33]

Perhaps the biggest disadvantage of conducting experiments in the field is that they are often challenging to implement. In contrast to the lab, where researchers can make unilateral decisions about what treatments to deploy, field experiments are often the product of coordination between researchers and those who actually carry out the interventions or furnish data on subjects' outcomes. Orr[34] and Gueron[35] offer helpful descriptions of how these partnerships are formed and nurtured over the course of a collaborative research project. Both authors stress the importance of building consensus about the use of random assignment. Research partners and funders sometimes balk at the idea of randomly allocating treatments, preferring instead to treat everyone or a hand-picked selection of subjects. The researcher must be prepared to formulate a palatable experimental design and to argue convincingly that the proposed use of random assignment is both feasible and ethical. The authors also stress that successful implementation of the agreed-upon experimental design—the allocation of subjects, the administration of treatments, and the measurement of outcomes—requires planning, pilot testing, and constant supervision.

Managing research collaboration with schools, police departments, retail firms, or political campaigns sounds difficult and often is. Nevertheless, field experimentation is a rapidly growing form of social science research, encompassing hundreds of

30 See Doleac and Stein 2010 for a study of racial discrimination by bidders on Internet auctions or Pager, Western, and Bonikowski 2009 for a study of labor market discrimination. We discuss discrimination experiments in Chapters 9 and 12.

31 In cases where surveys are used to assess outcomes, measurement may be unobtrusive in the more limited but nevertheless important sense that subjects are unaware that the survey aims to gauge the effects of the intervention.

32 Orchestrating return visits to the lab often presents logistical challenges, and failure to attract all subjects back to the lab may introduce bias (see Chapter 7).

33 See, for example, Gerber, Gimpel, Green, and Shaw 2011. See also the discussion of outcome measurement in Chapter 12.

34 Orr 1999, Chapter 5.

35 Gueron 2002.

studies on topics like education, crime, employment, savings, discrimination, charitable giving, conservation, and political participation.[36] The set of noteworthy and influential studies includes experiments of every possible description: small-scale interventions designed and implemented by researchers; collaborations between researchers and firms, schools, police agencies, or political campaigns; and massive government-funded studies of income taxes, health insurance, schooling, and public housing.[37]

Time and again, researchers overcome practical hurdles, and the boundaries of what is possible seem to be continually expanding. Consider, for example, research on how to promote government accountability. Until the 1990s, research in this domain was almost exclusively nonexperimental, but a series of pathbreaking studies have shown that one can use experiments to investigate the effects of government audits and community forums on accounting irregularities among public works programs,[38] the effects of grassroots monitoring efforts on the performance of legislators,[39] and the effects of information about constituents' preferences on legislators' roll call votes.[40] Field experiments are sometimes faulted for their inability to address big questions, such as the effects of culture, wars, or constitutions, but researchers have grown increasingly adept at designing experiments that test the effects of mechanisms that are thought to transmit the effects of the hard-to-manipulate variables.[41] Given the rapid pace of innovation, the potential for experimental inquiry remains an open question.

1.6 Naturally Occurring Experiments and Quasi-Experiments

Another way to expand the domain of what may be studied experimentally is to seize on *naturally occurring experiments*. Experimental research opportunities arise when interventions are assigned by a government or institution.[42] For example, the

36 Michalopoulos 2005; Green and Gerber 2008.

37 See, e.g., Robins 1985 on income taxes; Newhouse 1989 on health insurance; Krueger and Whitmore 2001 and U.S. Department of Health and Human Services 2010 on schooling. On public housing, see Sanbonmatsu et al. 2006; Harcourt and Ludwig 2006; and Kling, Liebman, and Katz 2007.

38 Olken 2007.

39 Humphreys and Weinstein 2010; Grose 2009.

40 Butler and Nickerson 2011.

41 Ludwig, Kling, and Mullainathan 2011; Card, Della Vigna, and Malmendier 2011.

42 Unfortunately, the term "natural experiment" is sometimes used quite loosely, encompassing not only naturally occurring randomized experiments but also any observational study in which the method of assignment is haphazard or inscrutable. We categorize studies that use near-random or arguably random assignment as quasi-experiments. For definitions of the term *natural experiment* that do not require random assignment, see Dunning 2012 and Shadish, Cook, and Campbell 2002, p. 17.

Vietnam draft lottery,[43] the random assignment of defendants to judges,[44] the random audit of local municipalities in Brazil,[45] lotteries that assign parents the opportunity to place their children in different public schools,[46] the assignment of Indian local governments to be headed by women or members of scheduled castes,[47] the allocation of visas to those seeking to immigrate,[48] and legislative lotteries to determine which representative will be allowed to propose legislation[49] are a few examples where randomization procedures have been employed by government, setting the stage for an experimental analysis. Researchers have also seized on natural experiments conducted by nongovernmental institutions. Universities, for example, occasionally randomize the pairing of roommates, allocation of instructors, and composition of tenure review committees.[50] Sports of all kinds use coin flips and lotteries to assign everything from the sequence of play to the colors worn by the contestants.[51] This list of naturally occurring experimental opportunities might also include revisiting random allocations conducted for other research purposes. A *downstream experiment* refers to a study whose intervention affects not only the proximal outcome of interest but, in so doing, potentially influences other outcomes as well (see Chapter 6). For example, a researcher might revisit an experiment that induced an increase in high school graduation rates in order to assess whether this randomly induced change in educational attainment in turn caused an increase in voter turnout.[52] In this book, we scarcely distinguish between field experiments and naturally occurring experiments, except to note that extra effort is sometimes required in order to verify that draft boards, court systems, or school districts implemented random assignment.

Quite different are *quasi-experiments*, in which near-random processes cause places, groups, or individuals to receive different treatments. Since the mid-1990s, a growing number of scholars have studied instances where institutional rules cause near-random treatment assignments to be allocated among those who fall just short of or just beyond a cutoff, creating a discontinuity. One of the most famous examples of this research design is a study of U.S. congressional districts in which one party's candidate narrowly wins a plurality of votes.[53] The small shift in votes that separates a narrow victory from a narrow defeat produces a treatment—winning the seat in the House of Representatives—that might be construed as random. One

43 Angrist 1991.
44 Kling 2006; Green and Winik 2010.
45 Ferraz and Finan 2008.
46 Hastings, Kane, Staiger, and Weinstein 2007.
47 Beaman et al. 2009; Chattopadhyay and Duflo 2004.
48 Gibson, McKenzie, and Stillman 2011.
49 Loewen, Koop, Settle, and Fowler 2010.
50 Sacerdote 2001; Carrell and West 2010; De Paola 2009; Zinovyeva and Bagues 2010.
51 Hill and Barton 2005; see also Rowe, Harris, and Roberts 2005 for a response to Hill and Barton.
52 Sondheimer and Green 2009.
53 Lee 2008.

could compare near-winners to near-losers in order to assess the effect of a narrow victory on the probability that the winning party wins reelection in the district two years later.

Because quasi-experiments do not involve an explicit random assignment procedure, the causal inferences they support are subject to greater uncertainty. Although the researcher may have good reason to believe that observations on opposite sides of an arbitrary threshold are comparable, there is always some risk that the observations may have "sorted" themselves so as to receive or avoid the treatment. Critics who have looked closely at the pool of congressional candidates who narrowly win or lose have pointed out that there appear to be systematic differences between near-winners and near-losers in terms of their political resources.[54]

The same concerns apply to a wide array of quasi-experiments that take weather patterns, natural disasters, colonial settlement patterns, national boundaries, election cycles, assassinations and so forth to be near-random "treatments." In the absence of random assignment, there is always some uncertainty about how nearly random these treatments are. Although these studies are similar in spirit to field experimentation insofar as they strive to illuminate causal effects in real-world settings, they fall outside the scope of this book because they rely on argumentation rather than randomization procedures. In order to present a single, coherent perspective on experimental design and analysis, this book confines its attention to randomized experiments.

1.7 Plan of the Book

This chapter has introduced a variety of important concepts without pausing for rigorous definitions or proofs. Chapter 2 delves more deeply into the properties of experiments, describing in detail the underlying assumptions that must be met for experiments to be informative. Chapter 3 introduces the concept of sampling variability, the statistical uncertainty introduced whenever subjects are randomly allocated to treatment and control groups. Chapter 4 focuses on how covariates, variables that are measured prior to the administration of the treatment, may be used in

54 Grimmer et al. 2011; Caughey and Sekhon 2011. In addition, regression discontinuity analyses often confront the following conundrum: the causal effect of the treatment is identified at the point of discontinuity, but data are sparse in the close vicinity of the boundary. One may expand the comparison to include observations farther from the boundary, but doing so jeopardizes the comparability of groups that do or do not receive the treatment. In an effort to correct for unmeasured differences between the groups, researchers typically use regression to control for trends on either side of the boundary, a method that introduces a variety of modeling decisions and attendant uncertainty. See Imbens and Lemieux 2008 and Green et al. 2009.

experimental design and analysis. Chapters 5 and 6 discuss the complications that arise when subjects are assigned one treatment but receive another. The so-called *noncompliance* or *failure-to-treat* problem is sufficiently common and conceptually challenging to warrant two chapters. Chapter 7 addresses the problem of attrition, or the failure to obtain outcome measurements for every subject. Because field experiments are frequently conducted in settings where subjects communicate, compare, or remember treatments, Chapter 8 considers the complications associated with interference between experimental units. Because researchers are often interested in learning about the conditions under which treatment effects are especially strong or weak, Chapter 9 discusses the detection of heterogeneous treatment effects. Chapter 10 considers the challenge of studying the causal pathways by which an experimental effect is transmitted. Chapter 11 discusses how one might draw generalizations that go beyond the average treatment effect observed in a particular sample and apply them to the average treatment effect in a broader population. The chapter provides a brief introduction to meta-analysis, a statistical technique that pools data from multiple experiments in order to summarize the findings of a research literature. Chapter 12 discusses a series of noteworthy experiments in order to highlight important principles introduced in previous chapters. Chapter 13 guides the reader through the composition of an experimental research report, providing a checklist of key aspects of any experiment that must be described in detail. Appendix A discusses regulations that apply to research involving human subjects. In order to encourage you to put the book's ideas to work, Appendix B suggests several experimental projects that involve low cost and minimal risk to human subjects.

SUGGESTED READINGS

Accessible introductions to experimental design in real-world settings can be found in Shadish, Cook, and Campbell 2002 and Torgerson and Torgerson 2008. For a discussion of the limitations of field experimentation, see Heckman and Smith 1995. Morgan and Winship (2007), Angrist and Pischke (2009), and Rosenbaum (2010) discuss the challenges of extracting causal inferences from nonexperimental data. Imbens and Lemieux (2008) provide a useful introduction to regression-discontinuity designs.

EXERCISES: CHAPTER 1

1. Core concepts:
   (a) What is an experiment, and how does it differ from an observational study?
   (b) What is "unobserved heterogeneity," and what are its consequences for the interpretation of correlations?
2. Would you classify the study described in the following abstract as a field experiment, a naturally occurring experiment, a quasi-experiment, or none of the above? Why?

   "This study seeks to estimate the health effects of sanitary drinking water among low-income villages in Guatemala. A random sample of all villages with fewer than 2,000 inhabitants was selected for analysis. Of the 250 villages sampled, 110 were found to have unsanitary drinking water. In these 110 villages, infant mortality rates were, on average, 25 deaths per 1,000 live births, as compared to 5 deaths per 1,000 live births in the 140 villages with sanitary drinking water. Unsanitary drinking water appears to be a major contributor to infant mortality."

3. Based on what you are able to infer from the following abstract, to what extent does the study described seem to fulfill the criteria for a field experiment?

   "We study the demand for household water connections in urban Morocco, and the effect of such connections on household welfare. In the northern city of Tangiers, among homeowners without a private connection to the city's water grid, a random subset was offered a simplified procedure to purchase a household connection on credit (at a zero percent interest rate). Take-up was high, at 69%. Because all households in our sample had access to the water grid through free public taps . . . household connections did not lead to any improvement in the quality of the water households consumed; and despite a significant increase in the quantity of water consumed, we find no change in the incidence of waterborne illnesses. Nevertheless, we find that households are willing to pay a substantial amount of money to have a private tap at home. Being connected generates important time gains, which are used for leisure and social activities, rather than productive activities."[55]

4. A parody appearing in the *British Medical Journal* questioned whether parachutes are in fact effective in preventing death when skydivers are presented with severe "gravitational challenge."[56] The authors point out that no randomized trials have assigned parachutes to skydivers. Why is it reasonable to believe that parachutes are effective even in the absence of randomized experiments that establish their efficacy?

55 Devoto et al. 2011.
56 Smith and Pell 2003.