Quasi-Experimentation

A GUIDE TO DESIGN AND ANALYSIS

CHARLES S. REICHARDT

Quasi-Experimentation

Methodology in the Social Sciences

David A. Kenny, Founding Editor

Todd D. Little, Series Editor

www.guilford.com/MSS

This series provides applied researchers and students with analysis and research design books that emphasize the use of methods to answer research questions. Rather than emphasizing statistical theory, each volume in the series illustrates when a technique should (and should not) be used and how the output from available software programs should (and should not) be interpreted. Common pitfalls as well as areas of further development are clearly articulated.

RECENT VOLUMES

CONFIRMATORY FACTOR ANALYSIS FOR APPLIED RESEARCH,
Second Edition
Timothy A. Brown

PRINCIPLES AND PRACTICE OF STRUCTURAL EQUATION MODELING,
Fourth Edition
Rex B. Kline

HYPOTHESIS TESTING AND MODEL SELECTION IN THE SOCIAL SCIENCES
David L. Weakliem

REGRESSION ANALYSIS AND LINEAR MODELS: Concepts, Applications,
and Implementation
Richard B. Darlington and Andrew F. Hayes

GROWTH MODELING: Structural Equation and Multilevel Modeling Approaches
Kevin J. Grimm, Nilam Ram, and Ryne Estabrook

PSYCHOMETRIC METHODS: Theory into Practice
Larry R. Price

INTRODUCTION TO MEDIATION, MODERATION, AND CONDITIONAL
PROCESS ANALYSIS: A Regression-Based Approach, Second Edition
Andrew F. Hayes

MEASUREMENT THEORY AND APPLICATIONS FOR THE SOCIAL SCIENCES
Deborah L. Bandalos

CONDUCTING PERSONAL NETWORK RESEARCH: A Practical Guide
Christopher McCarty, Miranda J. Lubbers, Raffaele Vacca, and José Luis Molina

QUASI-EXPERIMENTATION: A Guide to Design and Analysis
Charles S. Reichardt

Quasi-Experimentation

A Guide to Design and Analysis

Charles S. Reichardt

Series Editor's Note by Todd D. Little

THE GUILFORD PRESS
New York London

A Division of Guilford Publications, Inc.
370 Seventh Avenue, Suite 1200, New York, NY 10001
www.guilford.com

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data
Names: Reichardt, Charles S., author.
Title: Quasi-experimentation : a guide to design and analysis / Charles S. Reichardt.
Description: New York : Guilford Press, [2019] | Series: Methodology in the social sciences | Includes bibliographical references and index.
Identifiers: LCCN 2019017566| ISBN 9781462540204 (pbk.) | ISBN 9781462540259 (hardcover)
Subjects: LCSH: Social sciences—Experiments. | Social sciences—Methodology. | Experimental design.
Classification: LCC H62 .R4145 2019 | DDC 001.4/34—dc23
LC record available at *https://lccn.loc.gov/2019017566*

For Stefan, Grace, and Anne

Series Editor's Note

Research is all about drawing valid conclusions that inform policy and practice. The randomized clinical trial (RCT) has evolved as the gold standard for drawing causal inferences but it really isn't the golden chariot of valid inference. It's not fool's gold either—it's a sound design; but, thankfully, researchers do have other options, and sometimes these other options are better suited for a specific research question, particularly in field settings. Chip Reichardt brings you the wonderful world of valid and useful designs that, when properly implemented, provide accurate findings. His book is a delightful guide to the fundamental logic in this other world of inferential research designs—the quasi-experimental world.

As Reichardt indicates, the distinction between experimental and nonexperimental or quasi-experimental is more in the thoughtfulness with which the designs are implemented and in the proper application of the analytics that each design requires. Even RCTs can yield improper conclusions when they are degraded by factors such as selective attrition, local treatment effects, treatment noncompliance, variable treatment fidelity, and the like, particularly when implemented in field settings such as schools, clinics, and communities. Reichardt brings a thoughtful and practical discussion of all the issues you need to consider to demonstrate as best as possible the counterfactual that is a hallmark of accurate inference.

I like the word *verisimilitude*—the truthlike value of a study's results. When you take Reichardt's advice and implement his tips, your research will benefit by having the greatest extent of verisimilitude. In this delightfully penned book, Reichardt shares his vast state-of-the craft understanding for valid conclusions using all manner of inferential design. Theoretical approaches to inferential designs have matured considerably, particularly when modern missing-data treatments and best-practice statistical methods are employed. Having studied and written extensively on these designs, Reichardt

is at the top of the mountain when it comes to understanding and sharing his insights on these matters. But he does it so effortlessly and accessibly. This book is the kind you could incorporate into undergraduate curricula where a second course in design and statistics might be offered. For sure it is a "must" at the graduate level, and even seasoned researchers would benefit from the modernization that Reichardt brings to the inferential designs he covers.

Beyond the thoughtful insights, tips, and wisdom that Reichardt brings to the designs, his book is extra rich with pedagogical features. He is a very gifted educator and expertly guides you through the numbered equations using clear and simple language. He does the same when he guides you through the output from analysis derived from each of the designs he covers. Putting accessible words to numbers and core concepts is one of his super powers, which you will see throughout the book as well as in the glossary of key terms and ideas he compiled. His many and varied examples are engaging because they span many disciplines. They provide a comprehensive grounding in how the designs can be tailored to address critical questions with which we all can resonate.

Given that the type of research Reichardt covers here is fundamentally about social justice (identifying treatment effects as accurately as possible), if we follow his lead, our findings will change policy and practice to ultimately improve people's lives. Reichardt has given us this gift; I ask that you pay it forward by following his lead in the research you conduct. You will find his sage advice and guidance invaluable. As Reichardt says in the Preface, "Without knowing the varying effects of treatments, we cannot well know if our theories of behavior are correct or how to intervene to improve the human condition." As always, enjoy!

Todd D. Little
Society for Research in Child Development meeting
Baltimore, Maryland

Preface

Questions about cause and effect are ubiquitous. For example, we often ask questions such as the following: How effective is a new diet and exercise program? How likely is it that an innovative medical regimen will cure cancer? How much does an intensive manpower training program improve the prospects of the unemployed? How do such effects vary across different people, settings, times, and outcome measures? Without knowing the varying effects of treatments, we cannot well know if our theories of behavior are correct or how to intervene to improve the human condition. Quasi-experiments are designs frequently used to estimate such effects, and this book will show you how to use them for that purpose.

This volume explains the logic of both the design of quasi-experiments and the analysis of the data they produce to provide estimates of treatment effects that are as credible as can be obtained given the demanding constraints of research practice. Readers gain both a broad overview of quasi-experimentation and in-depth treatment of the details of design and analysis. The book brings together the insights of others that are widely scattered throughout the literature—along with a few insights of my own. Design and statistical techniques for a full coverage of quasi-experimentation are collected in an accessible format, in a single volume, for the first time.

Although the use of quasi-experiments to estimate the effects of treatments can be highly quantitative and statistical, you will need only a basic understanding of research methods and statistical inference, up through multiple regression, to understand the topics covered in this book. Even then, elementary statistical and methodological topics are reviewed when it would be helpful. All told, the book's presentation relies on common sense and intuition far more than on mathematical machinations. As a result, this book will make the material easier to understand than if you read the original literature on your own. My purpose is to illuminate the conceptual foundation of

quasi-experimentation so that you are well equipped to explore more technical literature for yourself.

While most writing on quasi-experimentation focuses on a few prototypical research designs, this book covers a wider range of design options than is available elsewhere. Included among those are research designs that remove bias from estimates of treatment effects. With an understanding of the complete typology of design options, you will no longer need to choose among a few prototypical quasi-experiments but can craft a unique design to suit your specific research needs. Designing a study to estimate treatment effects is fundamentally a process of pattern matching. I provide examples from diverse fields of the means to create the detailed patterns that make pattern matching most effective.

ORGANIZATION AND COVERAGE

I begin with a general overview of quasi-experimentation, and then, in Chapter 2, I define a treatment effect and the hurdles over which one must leap to draw credible causal inferences. Chapter 3 explains that the effect of a treatment is a function of five size-of-effect factors: the treatment or cause, the participants in the study, the times at which the treatments are implemented and effects assessed, the settings in which the treatments are implemented and outcomes assessed, and the outcome measures upon which the effects of the treatment are estimated. Also included is a simplified perspective on threats to validity. Chapter 4 introduces randomized experiments because they serve as a benchmark with which to compare quasi-experiments.

Chapter 5 begins the discussion of alternatives to randomized experiments by starting with a design that is not even quasi-experimental because it lacks an explicit comparison of treatment conditions. Chapters 6–9 present four prototypical quasi-experiments: the pretest–posttest design, the nonequivalent group design, the regression discontinuity design, and the interrupted time-series design. The threats to internal validity in each design that can bias the estimate of a treatment are described, along with the methods for coping with these threats, including both simple and advanced statistical analyses.

Since researchers need to be able to creatively craft designs to best fit their specific research settings, Chapter 10 presents a typology of designs for estimating treatment effects that goes beyond the prototypical designs in Chapters 6–9. Chapter 11 shows how each of the fundamental design types in the typology can be elaborated by adding one or more supplementary comparisons. The purpose of the additional comparisons is to rule out specific threats to internal validity. Such elaborated designs employ comparisons that differ in one of four ways: they involve different participants, times, settings, or outcome measures.

Chapter 12 describes and provides examples of unfocused design elaboration and explains how unfocused design elaboration can address the multiple threats to validity

that can be present. Chapter 12 also conceptualizes the process of estimating treatment effects as a task of pattern matching. The design typology presented in Chapter 10, together with the design elaborations described in Chapters 11 and 12, provide the tools by which researchers can well tailor a design pattern to fit the circumstances of the research setting. The book concludes in Chapter 13 with an examination of the underlying principles of good design and analysis.

Throughout, the book explicates the strengths and weaknesses of the many current approaches to the design and statistical analysis of data from quasi-experiments. Among many other topics, sensitivity analyses and advanced tactics for addressing inevitable problems, such as missing data and noncompliance with treatment assignment, are described. Advice and tips on the use of different design and analysis techniques, such as propensity score matching, instrumental variable approaches, and local regression techniques, as well as caveats about interpretation, are also provided. Detailed examples from diverse disciplinary fields illustrate the techniques, and each mathematical equation is translated into words.

Whether you are a graduate student or a seasoned researcher, you will find herein the easiest to understand, most up-to-date, and most comprehensive coverage of modern approaches to quasi-experimentation. With these tools, you will be well able to estimate the effects of treatments in field settings across the range of the social and behavioral sciences.

ACKNOWLEDGMENTS

My thinking about quasi-experimentation has been greatly influenced by interactions with many people over the years—especially Bob Boruch, Don Campbell, Tom Cook, Harry Gollob, Gary Henry, Mel Mark, Will Shadish, Ben Underwood, and Steve West. I benefited greatly from comments on the manuscript by several initially anonymous reviewers, three of whom are Manuel González Canché, Felix J. Thoemmes, and Meagan C. Arrastia-Chisholm. I am especially indebted to Steve West and Keith Widaman for their exceptionally careful reading of the manuscript and their detailed comments. C. Deborah Laughton provided invaluable guidance throughout the work on the volume. My sincere thanks to all.

Contents

1

Introduction

For as Hume pointed out, causation is never more than an inference; and any inference involves at some point the leap from what we see to what we can't see. Very well. It's the purpose of my Inquiry to shorten as much as humanly possible the distance over which I must leap. . . .

—BARTH (1967, p. 214)

Although purely descriptive research has an important role to play, we believe that the most interesting research in social science is about questions of cause and effect.

—ANGRIST AND PISCHKE (2009, p. 3)

The theory of quasi-experimentation . . . is one of the twentieth century's most influential methodological developments in the social sciences.

—SHADISH (2000, p. 13)

Overview

Researchers often assess the effects of treatments using either randomized experiments or quasi-experiments. The difference between the two types of designs lies in how treatment conditions are assigned to people (or other study units). If treatment conditions are assigned at random, the design is a randomized experiment; if treatment conditions are not assigned at random, the design is a quasi-experiment. Each design type has its place in the research enterprise, but quasi-experiments are particularly well suited to the demands of field settings. This volume explicates the logic underlying the design and analysis of quasi-experimentation, especially when they are implemented in field settings.

1.1 INTRODUCTION

We frequently ask questions about the effects of different treatments. Will attending college produce more income in the long run than attending a trade school? Which form of psychotherapy most effectively reduces depression? Which smoking cessation

program leads to the greatest reduction in smoking? Will this innovative reading program help close the gap in reading abilities between preschool children from high- and low-socioeconomic strata? How much does this criminal justice reform reduce recidivism among juveniles? And so on.

As these examples suggest, estimating the effects of treatments or interventions is of broad interest. Indeed, the task of estimating effects is of interest across the entire range of the social and behavioral sciences, including the fields of criminology, economics, education, medicine, political science, public health, public policy, psychology, social work, and sociology. In all of these fields, estimating effects is a mainstay in both basic research and applied research.

One of the primary tasks of **basic research** is testing theories—where a theory is tested by identifying its implications and making empirical observations to see if the implications hold true. And some of the most significant implications of theories involve predictions about the effects of treatments. So testing theories often involves estimating the effects of treatments. For example, consider tests of the theory of cognitive dissonance, a theory that specifies that when beliefs and behaviors are incongruent, people will change their beliefs to bring the beliefs into accord with the behaviors (Festinger, 1957). In a classic test of this theory, Aronson and Mills (1959) offered study participants membership in a discussion group if they would perform a disagreeable task (i.e., reading obscene material out loud). As predicted by the theory of cognitive dissonance, Aronson and Mills found that performing the disagreeable task increased the participants' liking for the discussion group. The idea was that cognitive dissonance would be aroused by performing the disagreeable task unless membership in the discussion group was viewed as desirable. So Aronson and Mills's research tested the theory of cognitive dissonance by estimating the effects of a treatment (performing a disagreeable task) on an outcome (liking for the group). In ways such as this, estimating the effects of treatments often plays a central role in theory testing.

Applied research is also concerned with estimating effects, though not always in the service of testing theories. Instead, applied research in the social sciences most often focuses on assessing the effects of programs intended to ameliorate social and behavioral problems. For example, in the service of finding treatments that can improve people's lives, economists have estimated the effects of job training programs on employment; educators have estimated the effects of class size on academic performance; and psychologists have assessed the effectiveness of innovative treatments for substance abuse. Applied researchers want to know what works better, for whom, under what conditions, and for how long—and this involves estimating effects (Boruch, Weisburd, Turner, Karpyn, & Littell, 2009).

The present volume is concerned with the task of estimating the effects of treatments whether for testing theories or ameliorating social problems. Of course, assessing treatment effects is not the only task in the social and behavioral sciences. Other central research tasks include discovering intriguing phenomena and devising theories to explain them. The fact that other tasks are important in no way diminishes the

importance of the task of assessing the effects of treatments. Without knowing the effects of treatments, we cannot know if our theories of behavior are correct. Nor can we know how to intervene to improve the human condition. The point is that we need to understand the effects of treatments if we are to have a good understanding of nature and function in the world. This book will show you how to estimate the effects of treatments using quasi-experiments.

1.2 THE DEFINITION OF QUASI-EXPERIMENT

As will be explained in greater detail in Chapter 2, estimating the effects of treatments requires drawing comparisons between different treatment conditions. Often, a comparison is drawn between a treatment condition and a no-treatment condition; but the comparison could instead be drawn between two alternative treatments. For example, a comparison could be drawn between an innovative treatment and the usual standard-of-care treatment.

Comparisons used to estimate the effects of treatments can be partitioned into two types: **randomized experiments** and **quasi-experiments** (Shadish, Cook, & Campbell, 2002). The difference has to do with how people (or other observational units such as classrooms, schools, or communities) are assigned to treatment conditions. In randomized experiments, study units are assigned to treatment conditions at random. Assigning treatment conditions at random means assigning treatments based on a coin flip, the roll of a die, the numbers in a computer-generated table of random numbers, or some equivalently random process. In quasi-experiments, units are assigned to treatment conditions in a nonrandom fashion, such as by administrative decision, self-selection, legislative mandate, or some other nonrandom process. For example, administrators might assign people to different treatment conditions based on their expectations of which treatment would be most effective for people with different characteristics. Alternatively, people might self-select treatments based on which treatment appears most desirable or most convenient.

I use the label **experiment** to refer to any randomized or quasi-experiment that estimates the effects of treatments. The label of experiment is sometimes used either more broadly or more narrowly. On the one hand, experiments sometimes are defined more broadly to include studies where no attempt is made to estimate effects. For example, a demonstration to show that an innovation can be successfully implemented might be called an experiment even if there is no attempt to assess the effects of the innovation. On the other hand, experiments are sometimes defined more narrowly, such as when the term is restricted to studies in which an experimenter actively and purposefully intervenes to implement a treatment (whether at random or not). Such usage would exclude what are called "natural" experiments which can arise, for example, when nature imposes a hurricane or earthquake or when ongoing social conventions, such as lotteries, serve to introduce interventions. And experiment is sometimes narrowly used

to be synonymous with randomized experiments. My usage is neither so broad nor so narrow. To me, an experiment is any attempt to estimate the effect of a treatment or an intervention using an empirical comparison. An experiment could involve an experimenter actively implementing a treatment, but the rubric of experiment also includes natural experiments. In any case, the exact usage of experiment is not as important as the distinction between randomized and quasi-experiments, both of which are experiments according to my nomenclature.

As another aside, randomized experiments are sometimes called "true" experiments. I avoid the label "true," for it suggests that alternatives to randomized experiments (i.e., quasi-experiments) are pejoratively false. Quasi-experiments are not the same as randomized experiments, but they are not false in any meaningful sense of the word. In addition, although they are often used interchangeably, I will not use the label randomized clinical trials (RCTs) instead of randomized experiment simply because RCT suggests to some readers an unnecessary restriction to medical or other health uses. For all intents and purposes, however, RCTs and randomized experiments mean the same thing.

Regardless of names and labels, the focus of this book is on quasi-experiments and especially on the logic and practice of quasi-experiments in field settings. However, this volume also considers randomized experiments. A researcher cannot fully appreciate quasi-experiments without reference to randomized experiments, so differences between the two are cited throughout the book. One of the first chapters is devoted to randomized experiments to provide a baseline with which to draw distinctions and lay the groundwork for the presentation of quasi-experiments.

1.3 WHY STUDY QUASI-EXPERIMENTS?

Randomized experiments are generally considered the gold standard for estimating effects, with quasi-experiments being relegated to second-class status (Boruch, 1997; Campbell & Stanley, 1966). So why do we need quasi-experiments when randomized experiments hold such an exalted position? One answer is that randomized experiments cannot always be implemented, much less implemented with integrity (West, Cham, & Liu, 2014). Implementing a randomized experiment would often be unethical. For example, it would be unethical to assess the effects of HIV by assigning people at random to be infected with the virus. Nor would it be ethical for researchers to randomly assign children to be physically abused or for couples to divorce. Even if we were to randomly assign children to be physically abused, we might not be sufficiently patient to wait a decade or two to assess the effects when the children become adults. Similarly, it is impractical, if not impossible, to assess the effects of such massive social interventions as recessions or wars by implementing them at random. Even when random assignment would be both ethical and physically possible, it can be difficult to convince both administrators and prospective participants in a study that randomized

experiments are desirable. Or funding agencies might require that investigators serve all of those most in need of a presumed ameliorative intervention, so that none of those most in need could be relegated to a presumed less effective comparison condition. People sometimes perceive random assignment to be unfair and therefore are unwilling to condone randomized experiments. Sometimes, too, data analyses are conducted after the fact when a randomized experiment has not been implemented or when a randomized experiment was implemented but became degraded into a quasi-experiment.

Even though randomized experiments can be superior to quasi-experiments in theory, they are not always superior in practice. Strict controls that can often be imposed in laboratory settings can enable randomized experiments to be implemented with high fidelity. But field settings often do not permit the same degree of control as in the laboratory; as a result, randomized experiments can become degraded when implemented in the field. For example, randomized experiments can be degraded when participants fail to comply with the assigned treatment conditions (which is called **noncompliance**) or drop out of the study differentially across treatment conditions (which is called differential **attrition**). Under some conditions, quasi-experiments can be superior to such corrupted randomized experiments. That is, it can sometimes be better to implement a planned quasi-experiment than to salvage a randomized experiment that has been corrupted in unplanned and uncontrolled ways.

Finally, even though randomized experiments are often superior to quasi-experiments, they are not perfect. Even well-implemented randomized experiments have weaknesses as well as strengths. Their strengths and weaknesses often complement the strengths and weaknesses of quasi-experiments. Science benefits from an accumulation of results across a variety of studies. Results accumulate best when the methods used to create knowledge are varied and complementary in their strengths and weaknesses (Cook, 1985; Mark & Reichardt, 2004; Rosenbaum, 2015b; Shadish, Cook, & Houts, 1986). Hence, the results of randomized experiments combined with quasi-experiments can be more credible than the results of randomized experiments alone (Boruch, 1975; Denis, 1990). Bloom (2005a, p. 15) expressed the idea in applied terms: "combining experimental and nonexperimental statistical methods is the most promising way to accumulate knowledge that can inform efforts to improve social interventions." Because randomized experiments are more often degraded in field settings than in the laboratory, quasi-experiments often best complement randomized experiments in field settings. Indeed, the strengths of the one particularly well offset the weaknesses of the other in field settings.

Randomized experiments are generally preferred to quasi-experiments in the laboratory and they are relatively more difficult to implement in the field than are quasi-experiments. For these reasons, quasi-experiments are more common in the field than in the laboratory (Cook & Shadish, 1994; Shadish & Cook, 2009), and the examples in this book are drawn mostly from the use of quasi-experiments in field settings. Nonetheless, the theory of quasi-experimentation is the same in both field and laboratory settings. What follows applies to both equally.

1.4 OVERVIEW OF THE VOLUME

This volume explains the logic of the design of quasi-experiments and the analysis of the data they produce to provide the most credible estimates of treatment effects that can be obtained under the many demanding constraints of research practice. The volume brings together the insights of others that are widely scattered throughout the literature. In addition, a few of my own insights are added that come from various locations in the literature. In this way, this work provides a compendium of material that would take substantial effort to assemble otherwise. In many cases, this volume makes this material easier to understand than if you read the original literature on your own. The purpose of this book is to provide accessible and helpful guidance to practitioners and methodologists alike.

Although estimating the size of effects can be highly quantitative and statistical, the presentation is directed to those who have no more than a basic understanding of statistical inference and the statistical procedure of multiple regression (and have taken an undergraduate course in research methods). Even then, elementary statistical and methodological topics are reviewed when it would be helpful. All told, the presentation relies on common sense and intuition far more than on mathematical machinations. I emphasize the conceptual foundation of quasi-experimentation so that readers will be well equipped to explore the more technical literature for themselves. When I have a choice to cite either a more or a less technical reference, I favor the less technical reference for its ease of understanding.

When I describe statistical procedures, please keep in mind that almost always multiple analysis strategies can be applied to data from any given quasi-experimental design (and new approaches are seemingly being developed every day). It would be impossible to present all the current possibilities, much less anticipate future developments. My purpose is to present the most common and basic analyses that will provide readers the framework they will need to understand both alternative variations and more sophisticated techniques—as well as future advances.

Chapter 2 defines a treatment effect and provides the background and conventions that undergird the ensuing presentation. A **treatment effect** is defined as a counterfactual difference between potential outcomes. Such a comparison is impossible to obtain in practice, so the definition of a treatment effect establishes the hurdles over which one must leap to draw credible causal inferences.

Chapter 3 explains that the effect of a treatment is a function of five **size-of-effect factors:** the treatment or cause; the participants in the study; the times at which the treatments are implemented and effects are assessed; the settings in which the treatments are implemented and outcomes assessed; and the outcome measures used to estimate the effects of the treatment. These five size-of-effect factors play a prominent role in distinguishing among types of quasi-experiments and in supplementing quasi-experimental designs to better withstand **threats to validity,** especially threats to internal validity (which are alternative explanations for obtained results). The five

size-of-effect factors are both the fundamental elements of an effect size and the fundamental components in the design of comparisons to estimate effects. Chapter 3 also introduces four types of validity and the notion of threats to validity. **Construct validity** is concerned with correctly labeling the five size-of-effect factors in a causal relationship. **Internal validity** is a special case of construct validity and is concerned with influences that are confounded with treatment assignment. **Statistical conclusion validity** addresses two questions: (1) Is the degree of uncertainty that exists in the estimate of a treatment effect correctly represented? and (2) Is that degree of uncertainty sufficiently small (i.e., is the estimate of the treatment effect sufficiently precise, and is a test of statistical significance sufficiently powerful?)? **External validity** concerns the generalizability of the study results.

Chapter 4 introduces randomized experiments because they serve as a benchmark with which to compare quasi-experiments and because randomized experiments can become degraded into quasi-experiments. As already noted, noncompliance to treatment assignment and differential attrition from treatment conditions can degrade randomized experiments. The means of coping with both types of degradation are presented (which can be considered part of the theory of quasi-experimentation). **Selection differences** are also addressed as a threat to internal validity.

Chapter 5 switches gears. While Chapter 4 concerns what is often said to be the gold standard of causal research design (i.e., the randomized experiment), Chapter 5 begins the discussion of alternatives to randomized experiments by starting with a design that is not even experimental because it does not entail an explicit comparison of treatment conditions. It is important to consider such a **pre-experimental design** because, despite its weaknesses, it is still used even though it is usually not considered acceptable research practice.

Chapters 6–9 present four prototypical quasi-experiments. In general, the presentation proceeds from the least to the most credible designs. Chapter 6 introduces the **pretest–posttest design** which is susceptible to what is often a debilitating range of biases, although the design can be used to good effect under limited circumstances. Chapter 7 introduces the **nonequivalent group design,** which is one of the most widely used quasi-experiments. Chapters 8 and 9 introduce the **regression discontinuity design** and the **interrupted time-series design**, respectively, which are the quasi-experiments that tend to produce the most credible results, though they are often the most demanding to implement. The threats to internal validity that most commonly arise in each design are described, and methods for coping with these threats (including statistical analyses) are explicated.

Although the designs presented in Chapters 6–9 are prototypical quasi-experimental designs, they do not cover the entire terrain of quasi-experiments. Because researchers need to be able to creatively craft designs to best fit their specific research settings, they need to know the full range of design options. Toward this end, Chapter 10 presents a typology of designs for estimating treatment effects that goes beyond the prototypical designs described in Chapters 6–9. The typology of designs distinguishes between

randomized experiments and quasi-experiments, as well as between two types of quasi-experiments: those where treatment assignment is determined according to a quantitative variable (e.g., the regression discontinuity and interrupted time-series designs) and those where treatment assignment is not so controlled (e.g., the pretest–posttest design and the nonequivalent group design). Cross-cutting the types of assignment to treatment conditions are four types of units that can be assigned to treatment conditions. The different units that can be assigned to treatments are participants, times, settings, and outcome measures. Chapter 10 shows where the four prototypical quasi-experimental designs fall within the more general typology and notes how the logic of the design and the analysis of the four prototypical designs generalize to the other designs in the typology.

Chapter 11 expands upon the typology presented in Chapter 10, showing how each fundamental design type in the typology can be elaborated by adding one or more supplementary comparisons. The additional comparisons help rule out specific threats to internal validity. Such elaborated designs employ comparisons that differ in one of four ways: they involve different participants, times, settings, or outcome measures. The design typology presented in Chapter 10, together with the elaborations, provide the tools by which researchers can tailor designs to fit the circumstances of their research settings. All too often the literature implies that researchers are to take a prototypical quasi-experimental design off the shelf, so to speak. That is, researchers are too frequently led to believe that the four prototypical quasi-experiments are the only choices available. The typology of designs in Chapter 10 and the elaborations in Chapter 11 provide a broader range of design options than is generally recognized. The task in designing a study is not to choose from among just four prototypes but to craft one's own design by selecting components from among the complete range of design options (Rosenbaum, 2015b, 2017; Shadish & Cook, 1999).

Chapter 12 distinguishes between focused and unfocused design elaborations. In **focused design elaborations,** separate estimates are used to address a shared threat to validity. Such focused design elaborations are explicated in Chapter 11. In **unfocused design elaborations**, separate estimates of the treatment effect are subject to different threats to validity. Chapter 12 provides examples of unfocused design elaboration and explains how unfocused design elaboration addresses the multiple threats to validity that are present. Chapter 12 also conceptualizes the process of estimating treatment effects as a task of **pattern matching**. To estimate a treatment effect, the researcher must collect data wherein the treatment is predicted to result in certain patterns of outcomes (should a treatment effect be present), while threats to validity are predicted to result in alternative patterns of outcomes (should they be present). The researcher then compares the predicted patterns to the data that are obtained. To the extent that the pattern predicted by the treatment fits the data better than the pattern predicted by threats to validity, the treatment is declared the winner and a treatment effect is plausible. Often, the best patterns for distinguishing treatment effects from the effects of threats to validity are complex. In quasi-experimentation, complex patterns are obtained by both

focused and unfocused design elaboration. Chapter 12 illustrates the benefits of complex patterns and accompanying complex designs in the context of unfocused design elaborations. Complex patterns can often be created by combining treatment comparisons. Therefore, a treatment comparison that is relatively weak if implemented on its own might nonetheless add substantially to the credibility of results when combined with other comparisons.

Chapter 13 concludes the presentation by describing underlying principles of good design and analysis. These principles include the following. Threats to internal validity are best ruled out using design features rather than statistical analyses. When in doubt about which underlying assumptions are correct, use multiple statistical analyses based on a range of plausible assumptions. **Treatment effect interactions,** and not just average treatment effects, should be assessed. Knowledge is best accumulated by critically combining results from a variety of perspectives and studies.

This volume also includes a glossary containing definitions of technical terms. Terms included in the glossary are bold-faced the first time they appear in the body of the text.

1.5 CONCLUSIONS

Both randomized and quasi-experiments estimate the effects of treatments. Some commentators have opined that only randomized experiments can satisfactorily fulfill that purpose. Such commentators fail to recognize both the frequent limitations of randomized experiments and the potential benefits of quasi-experiments. While randomized experiments are to be preferred to quasi-experiments in many instances, randomized experiments cannot always be implemented, especially in field settings—in which case quasi-experiments are the only option. Even when randomized experiments can be implemented, there can be benefits to quasi-experiments. Hence, researchers need to understand how to conduct quasi-experiments if they are to be able to estimate treatment effects when confronted with the diverse array of research needs and circumstances.

1.6 SUGGESTED READING

Campbell, D. T. (1969b). Reforms as experiments. *American Psychologist, 24,* 409–429.
—A classic call to action for using experimental methods to determine which social programs best ameliorate social problems.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology, 60,* 607–629.
—Presents a history of field experiments as used to ameliorate social problems.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
—A classic, must-read text on quasi-experimentation. The present volume provides more up-to-date coverage of quasi-experimentation than Shadish, Cook, and Campbell (2002). But if you want to know more about quasi-experimentation (especially experimental design) after reading the present volume, read Shadish et al.

West, S. G., Cham, H., & Liu, Y. (2014). Causal inference and generalizations in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 49–80). New York: Cambridge University Press.
—Also provides an insightful overview of randomized and quasi-experiments as implemented in the field.

2

Cause and Effect

The fact that causal inferences are made with considerable risk of error does not, of course, mean that they should not be made at all.
—BLALOCK (1964, p. 5)

In the past two decades, statisticians and econometricians have adopted a common conceptual framework for thinking about the estimation of causal effects—the counterfactual account of causality.
—WINSHIP AND MORGAN (1999, p. 661)

My attitude is that it is critical to define quantities carefully before trying to estimate them.
—RUBIN (2010, p. 40)

Overview

A treatment effect is the difference in outcomes between what happened after a treatment is implemented and what would have happened had a comparison condition been implemented instead, assuming everything else had been the same. Unfortunately, such a difference is impossible to obtain in practice because not everything else, besides the treatment and comparison conditions, can be held the same. Confounds are anything that cannot be held the same, and the presence of confounds can bias the estimate of a treatment effect. Violations of the **stable-unit-treatment-value assumption (SUTVA)** can also bias the estimate of a treatment effect. The present chapter explains how to avoid violations of SUTVA. Subsequent chapters explain how to cope with confounds due to threats to internal validity.

2.1 INTRODUCTION

The purpose of both randomized and quasi-experiments is to estimate the size of the effect of a treatment. Before we can undertake that task, we need to know what the effect of a treatment is. That is, to know how to estimate its size, we need to know how the effect of a treatment is defined.

The effect of a treatment is a difference because, as Holland (1986, p. 946) noted, "the effect of a cause is always relative to another cause" (also see Cook, 2005). What Holland meant is that an effect always involves a comparison between two treatment conditions—which I will call the treatment (or experimental) condition and the comparison condition. As a result, it is necessary to define a treatment effect in terms of a difference between a treatment and a comparison condition.

The effect of a treatment condition as compared to a specific comparison condition is the difference in outcomes between what happens after the treatment condition is implemented and what would have happened if the comparison condition had been implemented instead of the treatment condition, but everything else had been the same (Reichardt, 2006). For example, the effect on my headache pain of taking a couple of aspirin (as compared to taking a placebo) is defined as the difference between (1) the severity of my headache pain after taking the aspirin and (2) the severity of my headache pain if I had taken a placebo instead of the aspirin, but everything else had been the same. Suppose I had taken a couple of aspirin an hour ago and the headache I had then is now gone. Further suppose that if I had taken a placebo an hour ago rather than the aspirin, but everything else, including my headache, had been the same, I would still have the headache now. Then an effect of taking the aspirin (as compared to taking a placebo) is that my headache went away an hour later.

To be perfectly clear, let me restate the definition of a treatment effect with reference to Figure 2.1. As shown in the figure, a treatment effect is defined with reference to two points in time: the time at which the treatment and comparison conditions are introduced (Time 1) and the time at which the outcomes of the treatment and comparison conditions are assessed (Time 2). The effect of a treatment condition (as compared to a specified comparison condition) is the difference between (1) the outcome that would have arisen at Time 2 if the treatment condition had been implemented at Time 1 and (2) the outcome that would have arisen at Time 2 if the comparison condition had been implemented at Time 1 instead of the treatment condition, but (3) everything else at Time 1 had been the same.

The treatment in the treatment condition could be of any type—either one dimensional or multifaceted. The comparison condition might be a control condition where no treatment at all is imposed. Or the comparison condition might consist of an alternative

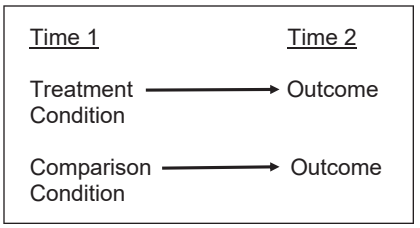


FIGURE 2.1. An effect is defined as the difference in outcomes at Time 2 when different treatment conditions had been implemented at Time 1, but everything else at Time 1 had been the same.

treatment condition such as the ordinary standard of care or business as usual. Or the comparison condition might consist of a placebo treatment, among other options. The comparison condition simply means a condition other than the treatment condition. In other words, in the definition of an effect, no restrictions are imposed on the nature of the treatment or comparison conditions. Also note that, although it is necessary, by definition, to specify an effect as a difference between a treatment condition and a comparison condition, it is common practice to talk of the effect of a treatment rather than, more appropriately, the effect of a difference between treatment and comparison conditions. I will often follow that practice. That is, I will often refer to a treatment effect without specifying an explicit comparison condition. Nonetheless, a difference between treatment and comparison conditions is always implied when speaking of a treatment effect. The nature of both the treatment and comparison conditions should always be spelled out in detail when either randomized or quasi-experiments are used in practice. Otherwise, a researcher will be estimating the effect of a difference between treatment and comparison conditions, but the audience will not know what that difference is.

2.2 PRACTICAL COMPARISONS AND CONFOUNDS

I will call the comparison of outcomes at Time 2 that defines a treatment effect the **ideal comparison** because it is impossible to obtain that comparison in practice. It is impossible because a researcher cannot both implement a treatment condition and *instead* implement a comparison condition at the same time with everything else being the same, as required by the definition of a treatment effect. For example, a researcher cannot both impose a treatment condition and impose a comparison condition instead of the treatment condition to the same person at the same time. I cannot both take a couple of aspirin and at the same time take a placebo instead of the aspirin. Nor can a researcher implement a treatment condition, see what happens, and then roll back time to implement a comparison condition under the same set of circumstances. I cannot take a couple of aspirin, see what happens, and then roll back time to take a placebo instead. As a result, the comparison of outcomes that defines a treatment effect is an impossible Platonic ideal. The impossibility of obtaining this ideal comparison is what Rubin (1978, p. 38) called the "fundamental problem facing inference for causal effects" (see also Holland, 1986).

An implication is that researchers cannot directly observe cause and effect; they can only infer it from imperfect data. Because the ideal comparison is impossible to obtain in practice, researchers can only infer the effect of a treatment using a practical comparison that differs from the ideal comparison. To estimate a treatment effect, a researcher must draw a comparison of outcomes. The comparison must be drawn between what happens after the treatment condition and after the comparison condition are implemented. However, such a comparison cannot be obtained where everything else had been held the same at Time 1 (as required in the ideal comparison). That

is, if a researcher is to draw a comparison between what happens after a treatment condition is implemented and what happens after a comparison condition is implemented (as is required to estimate a treatment effect), everything else cannot have been the same (as required in the ideal comparison). Something else, besides the treatment, must vary across the treatment and comparison conditions.

The something else, besides the treatment, that varies across the treatment and comparison conditions is called a **confound.** Different confounds will be present in different types of randomized experiments and quasi-experiments. But because something else, besides the treatment, must always vary across the treatment and comparison conditions, one or more confounds will always be present in any practical comparison used to estimate a treatment effect. For example, a researcher could estimate a treatment effect by comparing what happens when one group of participants receives the treatment condition and what happens when a different group of participants receives the comparison condition. But in that case, everything else is not the same because the participants who receive the treatment condition are not the same as the participants who receive the comparison condition. Thus, participants are confounded with the treatment conditions. (As will become clear in a subsequent chapter, such a confound is present even in **between-groups randomized experiments.**) Alternatively, a researcher could estimate a treatment effect by comparing what happens both before and after a treatment is implemented. In that case, however, the time before the treatment is implemented is not the same as the time after treatment is implemented, so time is confounded with the treatment conditions.

Because of the presence of one or more confounds, the estimate of the effect of a treatment in a practical comparison can be incorrect. Either part, or all, of any observed outcome difference in a practical comparison might be due to a confound that is present rather than to the difference in treatments received. For example, if a researcher compares the performance of a group of people who receive a treatment condition with the performance of a different group of people who receive the comparison condition, any observed difference might be due not to the difference in treatments but to initial differences between the two groups of people. Alternatively, if I compare the pain from my headache on Monday morning before I take a couple of aspirin to my pain on Monday afternoon after I take the two aspirin, any observed difference in headache pain may be due not to the aspirin but to differences in the time of day. The headache I had in the morning might have gone away by itself by the time the afternoon rolls around. Also note that a confound can either mask a treatment effect or masquerade as a treatment effect. A confound can either make an effective treatment look ineffective or make an ineffective treatment look effective.

Because of the presence of confounds, any practical comparison is only an approximation to the ideal but unobtainable comparison. Perhaps the approximation is a good one and perhaps not. How good an estimate of a treatment effect can be derived from a practical comparison depends on the quality of the approximation. In estimating a treatment effect, the task is to implement a sufficiently good approximation to the ideal

comparison so that, in reaching conclusions based on the obtained results, the presence of the inevitable confound or confounds can be properly taken into account. The primary purpose of this book is to explain how this task is to be accomplished—how to create an adequate substitute for the ideal but impossible comparison. In other words, in explicating both randomized experiments and quasi-experiments, my focus is on how these types of practical comparisons differ from the ideal comparison and how to take account of these inevitable differences.

2.3 THE COUNTERFACTUAL DEFINITION

My definition of a treatment effect is a **counterfactual definition**. The counterfactual label fits because the ideal comparison involves an outcome that is contrary to fact. That is, as noted earlier, it is impossible to implement both a treatment condition and a comparison condition instead, with everything else being the same, as required in the ideal comparison. If one outcome is obtained, the other cannot be obtained with everything else having been the same. The unobtained outcome is contrary to fact, hence the name "counterfactual."

The counterfactual definition, as given here, is the most common definition of a treatment effect. Some writers explicitly base their discussions of estimating treatment effects on a counterfactual definition like the one given here (Boruch, 1997; Holland, 1986; Reichardt, 2006; Reichardt & Mark, 1998; Rubin, 1974, 1977; Shadish et al., 2002; West, Biesanz, & Pitts, 2000). Many writers do not provide an explicit definition—either the one given here or any other. Nonetheless, the logic that underlies the practice of experimentation (which is a logic widely accepted across the social, behavioral, and statistical sciences) is compatible with the counterfactual definition given here. If you were to reverse-engineer the methods commonly used to estimate effects, you would find that they are undergirded by a counterfactual definition of a treatment effect.

In addition, the definition given here is fully compatible with the **Rubin causal model,** which is widely accepted in both statistics and the social and behavioral sciences. The Rubin causal model defines treatment effects as well as explicates the nature and role of mechanisms for selection into treatment conditions (Rubin, 2010; Shadish, 2010; Shadish & Sullivan, 2012; West & Thoemmes, 2010). The definition of a treatment effect in the Rubin causal model is the following (Holland 1986; Rubin, 1974, 2004, 2005; Splawa-Neyman, 1923, 1990). Let $Y_i(1)$ be the *i*th participant's response if that person were to receive the treatment condition and let $Y_i(0)$ be that same participant's response if that person were to receive the comparison condition instead of the treatment condition (but everything else were the same). Then the effect of the treatment (compared to the comparison condition) on the *i*th person is

$$Y_i(1) - Y_i(0) \tag{2.1}$$

And the effect of the treatment on a population of N participants (where $i = 1, 2, \ldots, N$) is the sum of the differences in Equation 2.1 across the N people:

$$\Sigma_{i=1,N} [Y_i(1) - Y_i(0)] \tag{2.2}$$

where $\Sigma_{i=1,N}$ indicates summation across the scores from i equals 1 to N. If the N participants in the summation consist of all the participants in the study, the estimate is called the **average treatment effect (ATE).** This is perhaps the most relevant estimate if the treatment is to be made mandatory for all potential participants (though the participants in the study might not represent all potential participants). If the summation is over only those participants who receive the treatment, the estimate is called the **average treatment effect on the treated (ATT)** or, equivalently, the **treatment-on-the-treated (TOT) effect.** This is perhaps the most relevant estimate when participation in the treatment will be voluntary (though who volunteers for the treatment might change once its effectiveness becomes known). The average treatment effect is a weighted combination of the average treatment effect on the treated (ATT) and the **average treatment effect on the untreated (ATU).** If the treatment effect is the same for everyone, all of these average treatment effects are the same. Other average treatment effects will be introduced later.

Note that either $Y_i(1)$ or $Y_i(0)$, but not both, can be observed. Therefore, the effect of a treatment (including the average effect of a treatment) cannot be directly observed. To estimate a treatment effect, a researcher must use a comparison other than this ideal, but unobtainable, one. In this way, the Rubin causal model leads to the same conclusions as the definition of a treatment effect given in Section 2.1.

Rubin (2005) prefers to say the Rubin causal model's definition of a treatment effect involves potential outcomes rather than a counterfactual outcome. His reasoning is the following. For an alternative outcome to be contrary to fact, a first outcome must have taken place. For example, outcome $Y_i(0)$ is contrary to fact only if outcome $Y_i(1)$ has been realized, or vice versa. If neither outcome has yet been realized, neither is contrary to fact. Rubin (2005) prefers to define a treatment effect even before either of the outcomes $Y_i(1)$ or $Y_i(0)$ has been realized. Nonetheless, the Rubin causal model becomes a counterfactual definition once one of the outcomes is instantiated. Except for different labels, however, the Rubin causal model is equivalent to the counterfactual definition given in Section 2.1. Both the potential outcomes and counterfactual labels are widely used. I prefer the counterfactual label only because it seems to me most descriptive given that, in any practical application, one of the outcomes is in fact realized so the other is counterfactual.

The critical point is the following. Despite occasional demurrals (e.g., Dawid, 2000; but see Rubin, 2000), there is widespread acceptance of a counterfactual definition of an effect, even if complete agreement on the best name for such a definition has not been reached. In my experience, when a definition of an effect is given explicitly, it is almost always a counterfactualist definition in agreement with the definition given in Section

2.1. And when a definition is not given, a counterfactualist definition usually provides the logical foundation upon which methods for estimating effects rest.

I don't believe there is an adequate alternative to a counterfactual definition of a treatment effect. Alternatives have been proposed, but they often fail because they are circular: they define the connection between cause and effect using the very notion of causality which they are supposed to be defining (cf. Mohr, 1995). Nonetheless, it is possible that an adequate alternative exists and could provide the basis for an alternative set of methods for estimating treatment effects. To avoid confusion that might thereby arise, researchers should be clear about the definition of a treatment effect they are using. My presentation of methods for assessing treatment effects rests explicitly on the foregoing, widely accepted, counterfactualist definition.

2.4 THE STABLE-UNIT-TREATMENT-VALUE ASSUMPTION

The stable-unit-treatment-value assumption (SUTVA) is a fundamental assumption for estimating effects that has been imposed in the context of the Rubin causal model (Rubin, 2004, 2008a, 2010). SUTVA states that, when exposed to a specified treatment, the outcome produced by a participant in a study will be the same (1) regardless of how the participants in the study were assigned to treatment conditions and (2) regardless of the treatments received by the other participants in the study. SUTVA also specifies that the potential outcomes in a study are well defined in the sense that there is only one potential outcome per participant under each treatment condition. Among other things, SUTVA implies that the amount of treatment is not limited in the sense that, if one treatment participant gets a full dose, another treatment participant must get less than a full dose because only limited amounts of the treatment are available.

SUTVA can be violated in a variety of ways. Choice can be therapeutic, while lack of choice can have less than salutary effects. So, for example, being assigned to a treatment condition at random might engender different responses than would arise if participants were either allowed to choose treatment conditions for themselves or selected for treatment conditions because of need or merit. Participants might comply better with treatment protocols when they choose their own treatments than when they are assigned to treatment conditions by others. Such results would violate SUTVA because outcomes would depend on the mechanisms by which participants were assigned to treatment conditions. Researchers need to appreciate how the mechanism of treatment assignment, and not just the treatments themselves, might affect outcomes. If the mechanism affects the outcomes, researchers need to limit their conclusions accordingly.

SUTVA could also be violated in a study of a vaccination for a communicable disease if giving the vaccine to one participant reduced the chance another participant would be infected by the disease. SUTVA could also be violated if a student receiving an educational intervention influenced the outcome of another participating student residing in the same classroom. Or SUTVA could be violated because of **externalities**

in a job training program where there are a limited number of jobs available, so that a treatment enabling one participant to get a job meant another participant was kept from getting a job.

Such violations of SUTVA can be minimized or avoided by ensuring that the participants in the study are physically separated from one another. For example, quarantining persons infected with a communicable disease could avoid cross contamination. Alternatively, violations of SUTVA can sometimes be avoided by enlarging the units of assignment and analysis. For example, if both members of a couple were enrolled in a study of an HIV vaccine, the unit of assignment to treatment conditions (and hence the unit of analysis) should be the couple rather than each person individually. Rather than assigning students within a classroom to treatments individually, entire classrooms could be assigned to treatment conditions as a whole—with the treatment effect estimated by comparing whole classrooms rather than individual students within classrooms. For a job training program, it might be that entire communities would have to be assigned to treatments rather than persons within communities to avoid externalities.

Violations of SUTVA can also arise owing to compensatory rivalry, resentful demoralization, diffusion or imitation of treatments, and compensatory equalization of treatments (Shadish et al., 2002). **Compensatory rivalry** arises when people who are assigned to the comparison condition perform better than they would have otherwise because they are aware that other people received a more desirable treatment. This source of bias is also called the John Henry effect after the mythical steel driver who exerted greater effort than normal because he was placed in competition with a steam-powered hammer. **Resentful demoralization** is the opposite of compensatory rivalry: people who are assigned to the comparison condition perform worse because they perceive that others are receiving a more desirable treatment, and so they become demoralized over their relative disadvantage. Examples of resentful demoralization are provided by Fetterman (1982) and Lam, Hartwell, and Jekel (1994). Compensatory rivalry and resentful demoralization might be reduced or avoided by offering treatments that are perceived to be equally desirable or by using a wait-list comparison condition wherein the treatment condition is promised to participants in the comparison condition after the study is completed, should the treatment prove beneficial.

Diffusion or imitation of treatments means the treatment given to participants in the treatment condition becomes available to participants in the comparison condition. Diffusion or imitation of treatments might occur, for example, in a school where teachers are the treatment providers. Some teachers (labeled experimental teachers) are enlisted and trained to provide an educational innovation to their class. The remaining teachers (labeled comparison teachers) are not enlisted to provide the innovation and are expected to conduct their class as usual. However, the comparison teachers might learn of the innovation from the experimental teachers and implement the program in their own classrooms. In this way, the innovation would diffuse from the experimental classrooms to the comparison classrooms. For example, Cook et al. (1999) reported that

diffusion of treatments occurred, at least to some extent, in three out of ten comparison schools, owing to informal interactions among program administrators.

Finally, **compensatory equalization of treatments** means, for example, that administrators provide extra resources to the participants in the comparison condition to compensate for the advantages provided by the experimental treatment. For example, a school principal might devote extra resources to teachers assigned to a comparison condition to compensate for the additional resources an educational innovation provides to the teachers in the treatment condition. In this way, the difference between the treatment and comparison conditions is reduced so that there is less difference between the two groups in educational outcomes.

Diffusion or imitation of treatments and compensatory equalization of treatments (and some of the other violations of SUTVA) arise because of the awareness that different treatment conditions are given to different participants. These violations of SUTVA can be reduced or avoided by removing awareness of different treatment conditions such as by limiting communication between treatment conditions, including physically separating the units (e.g., by assigning treatments to classrooms in different schools rather than to classrooms within the same school). Such violations of SUTVA can also be avoided, as already suggested, by assigning and assessing larger study units—for example, by comparing classrooms rather than students or by comparing schools rather than classrooms. Even then, however, researchers need to be vigilant. Biglan, Ary, and Wagenaar (2000, p. 35), for example, documented that diffusion of treatments to different communities occurred even when the communities were spread out over 500 miles, where the communities were "separated from other communities by large areas of very-low-population-density agricultural land," and despite efforts to avoid diffusion. Thus, researchers need to be vigilant in implementing treatment comparisons to produce the intended treatment differences.

2.5 THE CAUSAL QUESTION BEING ADDRESSED

Many questions about causes and effects can be asked. To avoid confusion (which is commonplace), it is important to be clear about the causal question under consideration in this book. In particular, it is important to distinguish between the two most fundamental questions about cause and effect because only one is the focus of experimentation (Dawid, 2000; Holland, 1986; Reichardt, 2011c; Smith, 2013). The two fundamental questions are:

The Cause Question: What is a cause of a given effect?
The Effect Question: What is an effect of a given cause?

The **Cause Question** arises when a police detective solves a crime, when a physician diagnoses an illness, when an economist assesses the roots of poverty, or when a

program evaluator seeks to determine why an intervention failed to produce desired results. In all these cases, an effect is specified, and it is the cause of the effect that is being sought. For example, a police detective is confronted with a crime and seeks to identify the perpetrator. Similarly, a physician is presented with symptoms of an illness and seeks to identify the disease that caused them.

In contrast, the **Effect Question** reverses the role of cause and effect in the investigation. Instead of specifying an effect and seeking a cause (as in the Cause Question), the Effect Question specifies a cause and seeks to know its effects. That is, the investigator has a cause in the form of an intervention or treatment in mind and wishes to know its consequences. The purpose of experimentation is to answer the Effect Question: hence the need to define the effect of a given cause (i.e., the effect of a treatment) as was done in Section 2.1. In contrast, the Cause Question would require a definition of a cause of a given effect instead of the definition of an effect of a given cause. Let me emphasize these points. It is the Effect Question (rather than the Cause Question) that is of primary concern in experimentation. The focus of experimentation is to determine an effect for a given cause (i.e., a given treatment), not the reverse—hence the focus on the definition of an effect for a given cause.

Although experimentation focuses on the Effect Question, both the Cause Question and the Effect Question play central roles in science as well as in everyday life. But it is critical to be clear about which question is being asked because the methods for answering the one are not the same as the methods for answering the other.

It is also important to be clear about the question under investigation because it is all too easy to confuse the two questions, especially when an investigation of the Cause Question, as it so often does, turns into an investigation of the Effect Question. What happens is the following. The Cause Question ("What is the cause of a given effect?") is posed and given a tentative answer. Then the Effect Question ("What is the effect of a given cause?") is asked as a means of testing the tentative answer to the Cause Question. For example, in a famous homicide, a New York City resident named Kitty Genovese was repeatedly stabbed in front of an apartment building. According to the classic telling of the story, the assault took place over the span of half an hour and was observed by numerous residents inside the apartment building. Yet, despite the many observers, no one called the police to report the assault. Two social psychologists, John Darley and Bibb Latané, sought to explain why no one called the police. This is the Cause Question—the lack of action to call the police is the given effect, and the question is "what is the cause?" Darley and Latané hypothesized the cause was a diffusion of responsibility among the numerous observers of the crime. According to this explanation, the observers assumed other people were witnessing the crime and so assumed someone else would report it to the police. To conduct a test of that hypothesized answer, Darley and Latané turned the Cause Question around into an Effect Question. They induced a diffusion of responsibility (the cause) to see if it resulted in the absence of bystander intervention (the effect). Darley and Latané (1970) conducted a

broad array of studies demonstrating that a diffusion of responsibility (the given cause) could indeed result in reduced bystander intervention (the resulting effect). In other words, Darley and Latané's hypothesized answer to a Cause Question was assessed by examining an Effect Question.

Or consider another example. John Snow is credited with determining the cause of cholera during a 19th-century epidemic of the disease in London (Snow, 1855; Tufte, 1997). Before Snow's discovery, it was thought that cholera was transmitted by air rather than, what is the case, bodily fluids. By examining the pattern of cholera (the effect) in an isolated part of the city, Snow correctly surmised that the outbreak was due to drinking from a well that contained contaminated water (the cause). That is, he was presented with an effect (the cholera outbreak) and came up with a purported cause (contaminated water from a well). To test his theory, he turned the question around. He intervened by removing the handle from the pump (an intervention or treatment) and found that cases of cholera thereafter decreased (an effect). In other words, Snow started with the Cause Question (what is the cause of cholera?), provided a tentative answer (a contaminated water supply), and then switched to the Effect Question (what is the effect of removing the pump handle on the subsequent outbreak of cholera?) to test his tentative answer.

In the two preceding examples, researchers began their inquiries by asking the Cause Question and found that this led them to the Effect Question. The reverse is also the case. Researchers often start with the Effect Question and find themselves led to the Cause Question. For example, when researchers estimate the effects of treatments (the Effect Question), they often find themselves asking what other causes (i.e., confounds) might have been present that could have accounted for an observed effect estimate (the Cause Question).

Given the intimate interplay between the two questions in practice, it is easy to see how they might be confused. Another way they can be (and often are) confused is when objections to the Cause Question are mistaken to be objections to the Effect Question. For example, objections to the Cause Question sometimes arise because of the well-known difficulties in unambiguously defining what it means for something to be a cause of a given effect (Brand, 1976, 1979; Cook & Campbell, 1979; Mackie, 1974). Sometimes problems such as the **problem of overdetermination** and the **problem of preemption** (Cook, Scriven, Coryn, & Evergreen, 2010; Scriven, 2008, 2009), which are problems for the Cause Question, are incorrectly taken to be problems for the Effect Question (Reichardt, 2011c). (For more on the criticisms of overdetermination and preemption, see the Appendix.)

Such difficulties sometimes lead commentators to call for abandonment of all causal research, including abandonment of research to answer the Effect Question (Carr-Hill, 1968; Russell, 1913; Travers, 1981; cf. Scriven, 1968). Neither criticism mentioned above, however, justifies a call to abandon the Effect Question. First, while it has proven difficult to define a cause for a given effect, it is not difficult to define an effect for a given

cause. Indeed, as shown in Section 2.3, there is general agreement about how to define an effect for a given cause. As a result, there is no reason to abandon the Effect Question because of definitional difficulties. Second, as shown in the Appendix, some criticisms of a counterfactual definition that apply to the Cause Question do not apply to the Effect Question. Such criticisms provide no justification for abandoning the Effect Question.

In my view, both the Cause Question and the Effect Question should be pursued. Criticisms of causal research should be clear about which causal question is under dispute. As just noted, there appears to be little current debate that a counterfactual definition of an effect is a reasonable definition (see Section 2.3). The definition that supports the Effect Question is firmly established even if there is some doubt about an adequate definition for the Cause Question (see the Appendix). But our inability to provide a completely satisfactory definition of a cause is not sufficient reason to abandon the Cause Question, much less the Effect Question—which is the focus of this volume.

2.6 CONVENTIONS

Most of the time in what follows, I will consider randomized and quasi-experiments that have only two conditions—a treatment condition and a comparison condition (where the treatment condition is also called the experimental condition.) This restriction is for the sake of simplicity alone and results in no loss of generality because the logic of the presentation can easily be expanded to more complex designs that compare more than two treatment conditions. To be clear, it will often be useful to compare more than two conditions. Researchers might compare two different innovations to a standard treatment. Or researchers might use a **factorial design** to compare all combinations of two treatments. For example, with two treatments (*A* and *B*) there would be four treatment conditions. There would be four groups that receive either (1) neither treatment *A* nor treatment *B,* (2) treatment *A* but not treatment *B,* (3) treatment *B* but not treatment *A,* or (4) both treatment *A* and treatment *B*. An obvious advantage of such designs is that it allows the researcher to assess treatment effect interactions whereby the effect of a treatment is either enhanced or diminished by the presence of another treatment. Another possibility involving more than two treatment conditions is a study involving dose–response comparisons. Here the multiple treatments consist of varying levels (doses) of a treatment. For example, in an assessment of a new medication, researchers might compare doses of 10 mg, 20 mg, 30 mg, and so on.

I will often write as if there is a single outcome measure, but multiple outcome measures are possible and often to be recommended. Multiple outcome measures can be used to assess how effects vary over time, or multiple outcome measures can be added at the same time to assess how effects vary across different measurement constructs. I will even explain how it can often be advantageous to include some outcome measures that are expected to show effects due to the treatment, along with some outcome

measures that are not expected to show effects due to the treatment. The purpose of such **nonequivalent dependent variables** is to remove biases in the estimate of treatment effects.

In a similar vein, I will often write as if there is only a single treatment effect. As noted, however, there might be many different effects across different outcome measures. Equally important, there could be different effects across different participants in the study—another type of treatment effect interaction, also called **moderator effects.** So just because an average treatment effect is present does not mean the effect is the same for all participants. It could be that the treatment has a positive effect on some participants and a negative effect on others. Or it could be that the treatment has an effect on some, but not all, participants. For example, to say that smoking causes lung cancer is not to say that everyone who smokes gets lung cancer. The conclusion that smoking causes lung cancer merely implies that smoking causes some people to get lung cancer. Understanding such interaction or moderator effects can be critical in applying the results of a study to different potential participants.

In addition, to say that a treatment causes an effect is not meant to imply that the given treatment is the only cause of that effect. For example, to say that smoking causes lung cancer is not meant to imply that smoking is the only cause of lung cancer. Nonetheless, smoking is a noteworthy cause of lung cancer in that it causes a significant proportion of people to get lung cancer who otherwise would not have contracted the disease. But there are other causes of lung cancer as well.

I will assume that outcome variables are continuous rather than categorical. Most of the literature on estimating effects with quasi-experiments imposes such a restriction. The reason is that categorical outcome variables introduce complications in statistical analysis that do not interfere with the logic of quasi-experimentation. When the outcome variable is categorical, researchers will have to change some of the details of the statistical analyses, but the underlying logic of the design and analysis remains the same. Some statistical procedures also require certain technical assumptions (such as normal distributions of the outcome variables conditional on the independent variables) if they are to produce perfectly correct confidence intervals and statistical significance tests. To focus on the underlying logic, however, I do not list those technical assumptions in what follows. If the assumptions do not hold, alternative procedures are possible. For example, the central limit theorem removes many problems due to non-normality when the sample is sufficiently large. But with small sample sizes, well-known alternative procedures can be employed to cope with non-normality (Wilcox, 1996).

For convenience, it may be easiest to think of participants in a study as individual people. But the participants could also be conglomerates or aggregates of people such as households, classrooms, schools, and communities. For example, participants were countries in a study by Acemoglu, Johnson, and Robinson (2001), which asked "whether countries that inherited more democratic institutions from their colonial rulers later

enjoyed higher economic growth as a consequence" (Angrist & Pischke, 2009, p. 4). The context will usually make it clear when the participants are groups rather than individual people. I'll tend to use "participant" when talking of individual persons, but I will sometimes use "unit" when talking of conglomerates of people.

I should also note that the **units of assignment to treatment conditions** need not be either people or conglomerates of people. As explained in Chapter 10, units of assignment can also be times, settings, and outcome measures, in addition to participants. I will sometimes refer only to participants because that is the most obvious unit, but the discussion might apply to other types of units as well. In any case, there is no need for confusion. The units of assignment to treatment conditions will be made clear by the context.

2.7 CONCLUSIONS

The counterfactual definition of a treatment effect is the most defensible one. Rubin's causal model defines a treatment effect in terms of potential outcomes, but potential outcomes lead to the counterfactual definition once either of the potential outcomes is realized. Unfortunately, the counterfactual comparison that defines a treatment effect is unobtainable in practice. Not even a randomized experiment produces the ideal counterfactual comparison. Any comparison that is possible in practice can be biased because of confounds and violations of SUTVA. Researchers must be vigilant in identifying and coping with both sources of bias.

2.8 SUGGESTED READING

Cook, T. D., Scriven, M., Coryn, C. L. S., & Evergreen, S. D. H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31,* 105–117.

—Presents an exchange of diverging views about assumptions that underlie the methods of assessing treatment effects.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–970.

—Discusses the problems of causal inference in the context of the Rubin causal model.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics, 29,* 343–367.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331.

—Explicate the Rubin causal model and SUTVA.

In a special issue of the journal *Psychological Methods,* the following articles contrast the Campbellian (e.g., Shadish et al., 2002) perspective on estimating effects with the perspective on estimating effects provided by the Rubin causal model:

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, unreliable measurement and mode of data analysis. *Psychological Methods, 15*(1), 56–68.

Imbens, G. W. (2010). An economist's perspective on Shadish (2010) and West and Thoemmes (2010). *Psychological Methods, 15*(1), 47–55.

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods, 15*(1), 38–46.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15*(1), 3–17.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15*(1), 18–37.

3

Threats to Validity

All scientific inquiry is subject to error, and it is far better to be aware of this, to study the sources in an attempt to reduce it, and to estimate the magnitude of such errors in our findings, than to be ignorant of the errors concealed in the data. One must not equate ignorance of error with lack of error.

—Hyman (1954, p. 4; cited in Rosnow & Rosenthal, 1997, p. 1)

From philosopher John Stuart Mill, [the Campbellian approach to causality] took the idea that identifying a causal relationship requires showing that cause preceded effect, that cause covaries with effect, and that alternative explanations for the relationship between cause and effect are implausible. . . . Humans are poor at making many kinds of causal judgments, prone to confirmation bias, blind to apparent falsifications, and lazy about both design and identifying alternative explanations.

—Shadish (2010, pp. 7, 8)

In any research application, the answer we get is always limited to the specific way the treatment is constructed, the specific way the outcomes are measured, the particular population that is studied, the particular settings in which the study takes place, and the specific time period in which the study takes place.

—Hallberg, Wing, Wong, and Cook (2013, p. 225)

Overview

The size of an effect is a function of five factors: the cause, participant, time, setting, and outcome measure. A threat to construct validity arises when one or more of these five size-of-effect factors is mislabeled. For example, the cause of an effect would be mislabeled (and hence a threat to construct validity would arise) if a researcher described an outcome as due to the active ingredients in a treatment, when the outcome was actually due to a placebo effect. Internal validity is a special case of construct validity. A threat to internal validity arises when the cause of an effect has been mislabeled because a (certain type of) confound is present. Statistical conclusion validity concerns the accuracy with which treatment effects are estimated. External validity has to do with the generalizability of treatment effect estimates.

3.1 INTRODUCTION

Campbell (1957) introduced the notion of threats to validity when he invented the distinction between internal and external validity. His 1957 article was later expanded into a book chapter by Campbell and Stanley (1963), which was subsequently published as a stand-alone volume (Campbell & Stanley, 1966). In these works, internal validity was defined as answering the question "Did in fact the experimental stimulus make some significant difference in this specific instance?" External validity was defined as answering the question "To what populations, settings, and variables can this effect be generalized?" A threat to validity is anything that jeopardizes the answers to either one of these questions. Campbell and Stanley (1966) listed eight threats to internal validity and four threats to external validity.

Cook and Campbell (1976, 1979) added two more types of validity to Campbell's original internal and external categorization. The two new types of validity are statistical conclusion validity and construct validity. The addition of two new types of validity meant that the number of threats to validity was expanded from 12 to 33. Shadish et al. (2002) maintained the four types of validity of Cook and Campbell (1979) but added four more threats to validity, bringing the total to 37. Shadish et al. (2002) also altered the definitions of some of the four types of validity. The definitions of both construct and external validity were expanded. The distinction between internal and construct validity was reworked, and some threats to internal validity were recategorized as threats to construct validity.

The Campbellian treatment of threats to validity has not gone uncriticized (e.g., see Shadish et al., 2002). For example, Reichardt (2011b) noted that Shadish et al.'s (2002) explication of validity is not always internally consistent. To take one instance, Shadish et al. (2002, p. 34) emphasized that "validity is a property of inferences" and that validity refers to the truth of an inference: "When we say something is valid, we make a judgment about the extent to which relevant evidence supports that inference as being true or correct." But some of the threats to validity that Shadish et al. (2002) listed are threats not to the truth of an inference but to the precision of an estimate of a treatment effect, which is not the same as truth. For example, some threats to validity do not challenge whether a confidence interval for the size of a treatment effect is correct but whether it is sufficiently narrow (i.e., the estimate of a treatment effect is sufficiently precise). But criticizing a confidence interval for not being sufficiently narrow is not the same as challenging whether the confidence interval corresponds to its given level of confidence—which matters for the interval to be valid.

The point here is that the Campbellian definitions of validity and threats to validity have varied over time, are sometimes incongruous, and are not always easy to differentiate. As Shadish et al. (2002, p. 468) acknowledged, "So although the new formulation in this book is definitely more systematic than its predecessors, we are unsure whether that systematization will ultimately resort in greater terminological clarity or confusion."

To make matters worse, different methodological theorists have proposed different definitions and categorizations of types of validity (Cronbach, 1982; Mark, 1986). To emphasize the limited scope of internal validity, Campbell (1986) even proposed a name change—suggesting internal validity be relabeled "local molar causal validity"—but this nomenclature never caught on. In addition, research practitioners use the Campbellian notions of validity in conflicting ways. For example, many researchers characterize designs as either valid or invalid, which conflicts with Shadish et al.'s (2002) insistence that validity applies only to inferences. Or consider external validity. A study with volunteers as participants would, according to most researchers, yield externally invalid results if the findings did not also apply to nonvolunteers (e.g., Brewer & Crano, 2014; Jackson & Cox, 2013). According to Shadish et al. (2002), however, if an inference about volunteers were true, it could not be invalid even if it did not generalize to nonvolunteers because, as noted, validity and truth are synonymous. As Shadish et al. (2002, p. 474) note, "We are acutely aware of, and modestly dismayed at, the many different usages of these validity labels that have developed over the years and of the risk that poses for terminological confusion—even though we are responsible for many of these variations ourselves."

To further terminological clarity rather than confusion and to cope with the changes in the Campbellian definitions over time and their growing complexity, I translate Shadish et al.'s (2002) nomenclature into simpler language with simpler meanings—language and meanings that attempt to retain the underlying notions of validity that are consistent across the different Campbellian terminological reincarnations but are arguably easier to understand. To my way of thinking, threats to validity address three questions: "Are the conclusions drawn from a study correct"; "Are the conclusions sufficiently precise (which is not the same as being correct)"; and "Do the conclusions answer the questions of interest"? Before I can get to a consideration of threats to validity, however, I must explain how the size of a treatment effect is a function of five size-of-effect factors.

3.2 THE SIZE OF AN EFFECT

The size of an effect is a function of five factors: the cause (*C*), participant (*P*), time (*T*), setting (*S*), and outcome measure (*O*). These factors are called the size-of-effect factors.

3.2.1 Cause

The size of an effect depends on the nature of the treatment. This should be clear from the definition I have given of a treatment effect (see Section 2.1). If you change the nature of the treatment, you can change the size of the effect. For example, the effect of one aspirin can be different from the effect of two aspirin. Even more extreme, the effect

of two aspirin can be different from the effect of a dose of morphine or a vitamin pill. Therefore, the size of an effect is a function of the treatment or cause (*C*) of the effect. As noted in Section 2.1, the specification of the treatment includes the specification of both the treatment and comparison conditions.

3.2.2 Participant

The participant or participants (*P*) are the entity or entities that receive the treatment and upon whom outcomes are assessed. It is the behavior of the participants that the treatment is meant to influence. Often, participants are individual people, but they can also be nonhuman species. Or the participants can be aggregates of people such as classrooms, schools, hospital wards, neighborhoods, cities, or states, to name a few possible conglomerates. Or, to go in the other direction, the participants can be conceptualized as parts of people such as brains or tissues. In any case, the size of an effect depends on the participants, whatever they are. That is, the effect of a treatment can be different for different participants. For example, aspirin can have a quite different effect on someone who is allergic to aspirin than on someone who is not allergic. Or the effect of the treatment can be different for participants who are cooperative compared to those who are uncooperative (Kirk, 2009). Therefore, the size of an effect is a function of both the participants (*P*) and the treatment or cause (*C*).

3.2.3 Time

Times (*T*) are the chronological (i.e., historical) times both at which the treatment is implemented and at which the outcome measures are collected (see Figure 2.1). The difference between these two times specifies a time lag, which is also part of the specification of times. The size of an effect depends on the chronological time at which the treatment is implemented and the lag between the time the treatment is implemented and its effects are assessed. That is, the size of a treatment effect can be different at different times and different time lags. For example, the effect of two aspirin depends on the time lag between taking the aspirin and measuring the effects. Two minutes is too short a time for aspirin to affect headache pain. In contrast, 1 to 2 hours after taking aspirin are time lags likely to show the maximum effects of aspirin on headache pain. Several years might be required before the effects of aspirin on susceptibility to heart attacks are apparent. As these examples demonstrate, the size of an effect is a function of time (*T*).

3.2.4 Setting

The setting (*S*) is the environment in which the treatment is received by the participants and the environment in which the effect of the treatment is assessed. Settings might be laboratories, hospitals, schools, community centers, or street corners, to name just a few

possibilities. The size of an effect depends on the settings (*S*) in which the treatment or cause is implemented and the effect is assessed. That is, the effect of a treatment can be different in different settings. For example, if you have a headache and take a couple of aspirin, you might feel more relief if you are relaxing at home in the bath than if you are presenting an important project at a stressful work meeting. Results can differ in a laboratory compared to the field. For example, a laboratory might induce evaluation apprehension, impose demand characteristics, or inspire hypothesis guessing differently than in a field setting (Orne, 1963, 1969; Rosenthal, 1966; Rosenthal & Rosnow, 1969; Rosnow & Rosenthal, 1997). In this way, the size of an effect is a function of the setting (*S*).

3.2.5 Outcome Measure

An effect can be assessed on any number of different outcome measures (*O*). For example, the effect of taking a couple of aspirin can be assessed on headache pain or on the prevention of heart attacks. The effects of an educational program can be assessed on both math and verbal abilities. The effects of a job training program can be assessed on income, the quality of jobs obtained, or lengths of subsequent employment. The size of an effect will depend on which outcome measure is assessed. That is, the effect of a treatment can be different on different outcome measures. For example, aspirin might have substantial effects on headaches and prevention of heart attack but very little effect on obsessive-compulsive behaviors. In this way, the size of an effect is a function of the outcome measure (*O*).

3.2.6 The Causal Function

The five size-of-effect factors—cause (*C*), participant (*P*), time (*T*), setting (*S*), and outcome measure (*O*)—determine the size of an effect. This can be seen by returning to Figure 2.1. The two treatment conditions, the participants, the initial setting, and the initial time determine the two states of the world at Time 1 in Figure 2.1. The time lag and the outcome measures then determine the two outcomes at Time 2 in Figure 2.1, and the two outcomes determine the size of the effect. In other words, the size of an effect is determined by the combination of the five size-of-effect factors.

I have shown that an effect size varies with the five size-of-effect factors. I have also shown that these factors are sufficient to specify an effect size. Therefore, I have shown that an effect size is a function of the five size-of-effect factors of the cause, participant, time, setting, and outcome measure (CPTSO). In shorthand, I'll write this relationship or **causal function** as

Effect size (ES) equals a function of Cause (*C*), Participant (*P*),
Time (*T*), Setting (*S*), and Outcome Measure (*O*).

Even shorter notation produces

$$ES = f(CPTSO) \tag{3.1}$$

The five size-of-effect factors reference five traditional questions that are expected to be answered, for example, in newspaper stories. The five factors are the how (cause), who (participant), when (time), where (setting), and what (outcome measure) of an event, or in this case, of an effect size. To properly and unambiguously label an ES, a researcher must label all five size-of-effect factors.

The size-of-effect factors play a prominent role in both the present chapter and in later chapters, especially in Chapters 10 and 11. Reichardt's (2006) **principle of parallelism** is based on the four size-of-effect factors of participant, time, setting, and outcome measure as described in Chapter 10. The size-of-effect factors also play prominent roles in the methodologies of others including Cronbach (1982), Judd and Kenny (1981), and Shadish et al. (2002). I can now return to threats to validity—with reference to the five size-of-effect factors.

3.3 CONSTRUCT VALIDITY

According to Shadish et al. (2002, p. 38), construct validity is "the validity of inferences about the higher order constructs that represent sampling particulars." My translation is that construct validity has to do with whether the five size-of-effect factors in a conclusion about an effect size are correctly labeled (Cook, Tang, & Diamond, 2014). As just noted, labeling an effect size unambiguously requires labeling each of the five size-of-effect factors. Mislabeling any of the five is a threat to construct validity. Consider the labeling (and mislabeling) of each of the five size-of-effect factors in turn.

3.3.1 Cause

A quasi-experiment assesses the effects of a cause or treatment such as a job training program to increase employment, an intervention to teach reading skills, a behavioral modification therapy for smoking cessation, and the like. The construct validity of the cause (*C*) has to do with whether the treatment and comparison conditions are properly labeled.

Might an intervention labeled cognitive psychotherapy be psychodynamic therapy instead, or might the effect of the therapy be due to nonspecific aspects of the treatment such as personal contact with the therapist rather than to the specifics of the therapy (Chambless & Hollon, 2012)? Might a researcher intend to manipulate feelings of social isolation by having participants in a study wait in a room alone (versus waiting with others), but might that manipulation produce boredom or cognitive rumination instead (Brewer & Crano, 2014)? Might an intervention labeled an induction of

cognitive dissonance really be an induction of self-perception (Bem, 1972)? Was the treatment implemented at full strength and with appropriate fidelity as either implied or stated explicitly by the researcher (Sechrest, West, Phillips, Redner, & Yeaton, 1979; Yeaton & Sechrest, 1981; Zvoch, 2009)? For example, might a treatment mistakenly be said to have no effect when it would have a substantial effect were it implemented with adequate strength and integrity? Or, conversely, might a treatment have a beneficial effect only when implemented at full strength, but a researcher concludes that the treatment has positive effects when implemented even in small amounts (McGinley, 1997)? Did the researcher accurately portray the delivery, receipt, and adherence to the treatment regimens? Did the treatment innovation diffuse from the treatment group to the comparison group so that the difference between the treatment and comparison conditions was not as specified? Is an observed effect the result of experimenter expectancies, demand characteristics, evaluation apprehension, social desirability, Hawthorne effects, or confirmation bias rather than the intended treatment? These questions ask if the treatment and comparison conditions are properly labeled. They are the types of questions raised by concern for the **construct validity of the cause.** Threats to construct validity are present to the extent that treatment and comparison conditions are not properly labeled in the conclusions that are reached.

To avoid the construct invalidity of the cause, researchers should make sure the treatment is delivered and received as intended, but even researchers' best efforts will often not be sufficient to ensure that the treatment and comparison conditions are implemented as planned. Researchers need to be fully aware of the nature of the treatment contrast in whatever form it was in fact instantiated (Shadish et al., 2002)—and take any infidelities in treatment implementation into account when reporting the nature of the treatments to the study's audiences.

Note that some threats to the construct validity of the cause are confounds, while others are not. As noted in Section 2.2, a confound is something that varies in a comparison along with the treatment conditions but is not part of the intended treatment. For example, consider a placebo effect. A placebo effect can vary with treatment conditions—being present in the treatment condition but not in the comparison condition that has no placebo control. Therefore, assuming the researcher is interested in the effects of the presumed active ingredients in a treatment and not in a placebo effect, a placebo effect can be a confound. As a result, what is labeled an effect of the active ingredients in a treatment might instead be the effect of the active ingredients together with the effect of the placebo. If so, the cause of the effect has been mislabeled, and a placebo effect is a threat to the construct validity of the cause. Cook (2015; Cook et al., 2014) provides another example of a confound that is a threat to the construct validity of the cause—where video conferencing in a study of bail hearings was mistaken for the true causal agent that was actually a compositional shift in the judges making the bail recommendations.

Not all threats to the construct validity of the cause are confounds. Consider the degree to which the implementation of a given treatment is correctly specified. For example, suppose a researcher incorrectly states that a treatment implementation

follows the specifications in a manual when it does not fully do so. Such a mislabeling is a threat to the construct validity of the cause but is more easily conceptualized as a simple misrepresentation of the true treatment difference rather than as a confound.

3.3.2 Participant

A quasi-experiment involves one or more participants who might be unemployed workers, first-grade students, people who have unsuccessfully tried to quit smoking on their own, and so on. Construct validity includes the concern for whether the participants in the study were labeled properly. If the participants in the study were volunteers, did the researcher make that clear to his or her audience, or would a reader be likely to assume that the participants were not such a restricted sample? If the participants were untrained nurse assistants, did the researcher label them as such, or did the researcher mislabel them as trained health care professionals? Were the participants "good subjects" as assumed, or were they uncooperative? These questions ask if the participants in a study have been properly labeled. They are the types of questions raised by concern for the **construct validity of the participants.**

For a concrete example, consider an instance where the participants were mistakenly labeled too broadly. According to Schachter (1982), the clinical literature contains claims that the majority of those who quit smoking end up relapsing. But this result is based on a select group of people who quit smoking rather than all people who quit smoking. The percentages are based only on smokers who quit after seeking professional treatment because they are unable to quit smoking on their own. The percentages are said to apply to all smokers who quit—even those able to quit on their own. As Schachter (1982, p. 437) explains, "Our view of the intractability of the addictive states has been molded largely by the self-selected, hard-core group of people who, unable or unwilling to help themselves, go to therapists for help, thereby becoming the only easily available subjects for studies of recidivism and addiction."

The point is this: the relapse rate for those who can quit smoking on their own might be much lower than the relapse rate for those who seek professional assistance to quit smoking. If so, the percentages of those relapsing that are reported in the literature apply only to the group of smokers who seek professional help but are incorrectly described as if they apply to all smokers. In other words, the population of people from which the results were derived, and to which they properly apply, is mislabeled.

3.3.3 Time

A quasi-experiment assesses the effect of a treatment at a specific time and at a specific time interval after the treatment is implemented. Construct validity asks if the researcher's descriptions of the time and time lag for a treatment effect are accurate. The problem is that researchers often fail to recognize temporal variation in the effectiveness of a treatment (Cole & Maxwell, 2003; Gollob & Reichardt, 1987, 1991; MacCallum

& Austin, 2000; Maxwell & Cole, 2007; Maxwell, Cole, & Mitchell, 2011; Reichardt, 2011a; Reichardt & Gollob, 1986). The result is that researchers often imply that treatment effects hold over a wider range of times or time lags than is the case. For example, concluding that a treatment has no effect can suggest that a treatment has no effect over all time lags. But an estimate of no effect might have been derived at only a short time interval—before the treatment had time to realize its effects. Conversely, some treatments (such as antibiotics, pesticides, and advertisements) can have substantial effects at one point in time, but continued implementation over time can lead to reduced effectiveness, which might not be accurately conveyed in research reports. Such issues raise questions about whether time and time lags have been properly labeled; they raise questions about the **construct validity of time.**

3.3.4 Setting

A quasi-experiment assesses the effect of a treatment in a specific setting. A concern for the construct validity of the setting raises the question of whether the researcher's description of the setting was adequate and accurate. Did the researcher provide enough information for a reader to appreciate the extent to which the research environment induced evaluation apprehension, imposed demand characteristics, or inspired hypothesis guessing? Did the researcher acknowledge the extent to which the setting produced compensatory rivalry or resentful demoralization?

For an example, consider how the effects of psychotherapy are often assessed in the laboratory but have been assumed to apply to more general clinical settings. Need for caution is evidenced by Weisz, Weiss, and Donenberg (1992), who found the effects of psychotherapy on children to be greater in the laboratory than in "real-life" mental health clinics (but see also Shadish et al., 1997; Shadish, Matt, Navarro, & Phillips, 2000). Perhaps job training programs are more effective in times of economic prosperity than during economic downturns, but such restrictions on applicability go unreported. The effectiveness of ACT and SAT preparation courses in helping students get accepted to college would likely decrease to the extent that colleges place less emphasis on such tests of student abilities. Or such courses might increase the chances of an individual getting accepted to college but not in a setting in which all college applicants take test preparation courses. These are examples where settings could easily be mislabeled. They raise questions about the **construct validity of the setting.**

3.3.5 Outcome Measure

A quasi-experiment assesses the effect of a treatment on one or more outcome measures such as earnings, academic ability, cigarettes smoked in the past week, and the like. Construct validity asks if outcomes were measured in a fashion that is valid and reliable. The bottom line is that construct validity asks if the constructs measured correspond to the constructs that are said to be measured. Such misalignment is an issue because

no outcome measure is a pure measure of a construct. Measures of an intended construct might be influenced, for example, by observer expectancies—effects that were unintended and could easily go unrecognized by researchers. Or a test of mathematics ability might assess reading ability as well. Yet researchers might conclude that a treatment improves mathematical ability when the increase in "math" scores is due solely to improvements in reading ability. Or consider that decisions made within groups (as opposed to decisions made by individuals) were once thought to cause a shift toward riskiness on the outcome measure (Bem, Wallach, & Kogan, 1965; Wallace & Kogan, 1965). This effect has been subsequently recharacterized as a shift in what participants perceive as the most socially desirable choice (Abelson, 1995). These are examples of issues concerning the **construct validity of the outcome measure.** They raise questions about whether outcome measures have been properly labeled.

3.3.6 Taking Account of Threats to Construct Validity

Threats to construct validity are avoided by making sure that the causes, participants, times, settings, and outcome measures that are instantiated are both intended and labeled. For example, researchers should ensure that untrained nurses' assistants are labeled as such, rather than as trained health care professionals. Or to avoid the threat of placebo effects, researchers could equalize expectations about the treatment and the success of the comparison condition by giving the participants in the comparison condition a placebo.

Other times, threats to construct validity can be addressed using multiple operationalizations. For example, consider once again the test of cognitive dissonance by Aronson and Mills (1959). To gain admission to a discussion group, college women were induced to read embarrassingly obscene material. Subsequently, these women showed greater liking for the discussion group. Presumably, reading the obscene material led to greater liking because of cognitive dissonance, but this greater liking for the discussion group could have been the result of sexual arousal produced by the obscene material. To rule out this threat to the construct validity of the cause, subsequent research induced cognitive dissonance, using alternative manipulations that did not occasion sexual arousal. In other words, the threat to construct validity of the cause was removed by using multiple operationalizations of the treatment. More will be said about multiple operationalizations and multiple converging studies in Chapters 12 and 13.

3.4 INTERNAL VALIDITY

According to Shadish et al. (2002, p. 38), internal validity is "the validity of inferences about whether the observed covariation between *A* (the presumed treatment) and *B* (the presumed outcome) reflects a causal relationship from *A* to *B* as those variables were manipulated or measured." My translation is the following. Internal validity is a special

case of construct validity. While construct validity concerns the labeling (or mislabeling) of any of the five size-of-effect factors, internal validity concerns the labeling (or mislabeling) of only the size-of-effect factor of the cause (*C*). Internal validity concerns only certain mislabelings of the cause—only confounds that would have been present in a comparison even if the treatment had not been implemented. Confounds (such as placebo effects) that are present only because the treatment had been implemented are threats to construct validity. This last distinction may seem needlessly complex, but it is simply the way Shadish et al. (2002) characterize internal validity. Note, however, that while my definition of internal validity agrees with theirs in most ways, Shadish et al. (2002) would not agree with my characterization that internal validity is a special case of construct validity. But my conceptualization is in keeping with Shadish et al.'s explication of threats to validity and with how the definition of construct validity has evolved in measurement theory to be the overarching rubric for all types of measurement validity (Messick, 1989, 1995).

For an example of a threat to internal validity, consider a **between-groups comparison.** In the ideal comparison, the same participants are both given the treatment condition and, instead, given the comparison condition at the same time, which is impossible to do in practice (see Sections 2.1 and 2.2). In a between-groups comparison, which is possible to obtain in practice, two different groups of participants are compared at the same time where one group of participants receives the treatment condition and the other group receives the comparison condition. In such a comparison, initial differences between the participants in the two groups are confounded with the treatment so that any observed outcome differences between the groups could be due either to the treatment differences or to initial differences in the composition of the participants in the two groups (even in randomized experiments; see Section 4.4). Such initial-difference confounds are threats to internal validity because they could arise even in the absence of a treatment implementation (i.e., even in the absence of the treatment group receiving the treatment).

Confounds classified as threats to interval validity can arise from each of four sources: participants, times, settings, and outcome measures. Let me explain. The ideal comparison that defines a treatment effect cannot be obtained in practice because everything besides the treatment conditions cannot be held the same. In any practical comparison, something else must vary along with the treatment conditions. What must vary along with the treatment conditions is one or more of the four size-of-effect factors of participants, times, settings, and outcome measures. The variations in these factors are confounds and hence are potential sources of threats to internal validity. Consider threats to internal validity due to each of the four sources in turn.

3.4.1 Participant

As described earlier, participants could differ across the treatment conditions producing a confound that is a threat to internal validity. In between-groups comparisons,

such differences are called selection differences. More will be said about such differences in subsequent chapters.

3.4.2 Time

Times could differ across the treatment groups, producing a confound that is a threat to internal validity. For example, Dooley (1995) and Snyder (1974) report a case where differences in time lags occurred when blood from the treatment condition was left out in the air longer than blood from the comparison condition, thereby causing a difference in outcomes. Times are inherently confounded with treatment conditions in the pretest–posttest design and the interrupted time-series designs, which are described in subsequent chapters.

3.4.3 Setting

Settings could differ across the treatment conditions producing a confound that is a threat to internal validity. For example, the setting for the treatment group might contain additional interventions or annoyances compared to the setting for the comparison group.

3.4.4 Outcome Measure

Finally, outcome measures could differ across the treatment groups, producing a confound that is a threat to internal validity. For example, outcome measures might not be collected in the same manner in the treatment and comparison conditions because of different observers or data collectors.

When seeking to identify potential confounds, consider differences between the treatment conditions in participants, times, settings, and outcome measures. Much more will be said about internal validity in subsequent chapters with regard to specific types of designs.

3.5 STATISTICAL CONCLUSION VALIDITY

According to Shadish et al. (2002, p. 38), statistical conclusion validity is the "validity of inferences about the correlation (covariation) between treatment and outcome." I interpret this statement as follows. An examination of the nine threats to statistical conclusion validity listed in Shadish et al. reveals that statistical conclusion validity addresses two questions: (1) Is the degree of uncertainty that exists in a treatment effect estimate correctly represented? and (2) Is that degree of uncertainty sufficiently small (i.e., is the estimate of the treatment effect sufficiently precise?). The degree of uncertainty that exists in an estimate of a treatment effect can be represented by a confidence interval or by the results of a statistical significance test. Mistakes can be made in constructing

a confidence interval, so that a confidence interval that is said to contain the treatment effect with 95% confidence might really contain the treatment effect with only 50% confidence. Or a statistical significance test that is said to have a Type I error rate of 5% might really have a Type I error rate of 20%. In addition, conducting statistical significance tests of multiple outcome measures raises the chances of at least one Type I error, and this increased risk might be overlooked (Schochet, 2008). In these cases, the degree of uncertainty that is present is misspecified. Such mistakes threaten the statistical conclusion validity of a researcher's conclusion about a treatment effect.

At the same time, it is possible that the degree of uncertainty in an estimate of a treatment effect is properly specified, but the degree of uncertainty is too great. For example, a researcher might correctly construct a 95% confidence interval, but that interval might be so wide that it reveals little about the size of the treatment effect. Or a researcher might correctly conduct a statistical significance test with a Type I error rate of 5%. The power of the statistical significance test might be so low, however, that even a large treatment effect is likely to go undetected (Cohen, 1988). Both imprecise confidence intervals and low-powered statistical significance tests also threaten the statistical conclusion validity of a researcher's conclusion about a treatment effect.

Some statistical procedures produce both more precise estimates of treatment effects and more powerful tests of statistical significance than others. For example, researchers can measure pretreatment differences among participants that could account for differences in outcomes and include such measures in the statistical model as **covariates** (more on this in later chapters). Even more classic ways to increase power and precision are to increase the sample size of participants and to use more reliable outcome measures. Another approach to increasing power and precision is to use more homogeneous samples of treatments, participants, times, and settings. Subsequent chapters describe statistical procedures that, under the right conditions, produce valid assessments of uncertainty and lead to powerful tests of statistical significance and precise estimates of treatment effects.

3.6 EXTERNAL VALIDITY

According to Shadish et al. (2002, p. 83), "external validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes." My translation is that external validity has to do with the intended purpose of the study and with the generalizability of the study results. To be more precise, external validity asks whether the researcher studied the five size-of-effect factors of interest and, if not, whether the results that were obtained generalize to the size-of-effect factors of interest. To the extent that the results of a study generalize to the causes, participants, times, settings, and outcome measures that are of interest, the results are externally valid. Next let us consider generalizations to each of the five size-of-effect factors.

3.6.1 Cause

An experiment assesses the effects of a treatment condition in relation to a comparison condition. Were the treatment and comparison conditions the ones the researcher was interested in, and, if not, do the results generalize to the treatment and comparison conditions of interest? Perhaps the researcher was not able to implement the treatment at full strength. If the lack-of-full-strength implementation is correctly acknowledged, the inference does not suffer from construct invalidity. The same cannot necessarily be said about external validity. In this example, external validity asks if the obtained results (even if correctly labeled) generalize to what would have been the case if the treatment had been at full strength, or would the effect of a full-strength treatment have been different? If the results do not generalize to the full-strength treatment conditions as desired, the inferences suffer from external invalidity.

3.6.2 Participant

An experiment involves one or more participants. Were the participants in the study the ones in which the researcher was most interested? Might the participants have been restricted to volunteers while the researcher wanted to assess effects for nonvolunteers? And if the participants had been restricted to volunteers, would the results nonetheless generalize to nonvolunteers as the researcher and other stakeholders desire? If the results were obtained only from volunteers, the results would have construct validity if they had been correctly labeled as such. External validity is different. It asks if the results from volunteers would generalize to nonvolunteers, as desired.

As just one illustration, Hallberg et al. (2013) note that a program with a small number of participants might not have the same effects when scaled up in size so that there was a greater number of participants. For example, the benefits of small class sizes might not be achieved on a larger scale because of an inadequate supply of qualified teachers. If stakeholders were interested in the effects of small class sizes on a larger scale, the results of the study might not have external validity because they do not generalize to the desired (large) population of participants.

3.6.3 Time

An experiment is conducted at a specific time and with a specific time lag between when the treatment conditions are implemented and outcomes are assessed. Was the time lag between treatment and outcomes the one of interest to the researcher, and, if not, do the results from one time lag generalize to the other? Perhaps the researcher wants to know if a treatment had a long-lasting effect but was able to study the effect only over a short period of time. Do the results from the short period of time well predict the results after a longer period? Might short-term effects not last, or might effects that do not appear immediately be apparent after long time lags (Cook, Gruder, Hennigan, &

Flay, 1979; Murnane & Willett, 2011)? Such questions concern the generalizability (i.e., the external validity) of the results.

3.6.4 Setting

An experiment is conducted in a given setting. Was the setting the one the researcher was interested in, and, if not, do the results generalize to the setting of interest? Perhaps the setting of the research study induced evaluation apprehension in the participants, and this could have affected the results. Do the results generalize to settings in which the researcher is most interested; namely, settings that do not induce evaluation apprehension? Such questions concern the external validity of the results.

3.6.5 Outcome Measure

An experiment assesses the effects of a treatment on one or more outcome measures. Were the outcome measures the ones the researcher was interested in, and, if not, do the results generalize to the outcome measures of interest? Perhaps the researcher was interested in math achievement, but the outcome measure assessed reading ability as well as math achievement. Do the results generalize to a pure measure of math achievement? As noted by Koretz (2008), outcome measures can be carefully tailored to fit the treatment so that the treatment has a large effect because it "teaches to the test." In such instances, similarly large effects might not be obtained on alternative, more global outcome measures. Conversely, some treatments that are found to have no effect in a measured domain might have positive effects on other relevant (but unmeasured) domains of outcomes (Murnane & Willett, 2011). In either case, the results from outcome measures in one domain might not generalize to results from outcome measures in other domains, as desired, which are concerns of external validity.

3.6.6 Achieving External Validity

A classic method for obtaining generalizable results is sampling at random—which allows the generalization of obtained results to the randomly sampled population. For example, to be able to generalize the results of a randomized experiment to a given population of participants, a researcher could randomly sample participants from that population as suggested by Draper (1995) and Kish (1987). Bloom (2008) cites three studies that randomly sampled participants (along with randomly assigning participants to treatment conditions). But random sampling of participants is difficult to accomplish, and even if it is accomplished, it provides results that are still limited to participants at one point in time. For such reasons, Campbell (1988, p. 324) argued that historically "there was gross overvaluing of, and financial investment in, external validity, in the sense of representative samples at the nationwide level." Randomly sampling participants does not provide for random sampling of treatments, times,

settings, and outcome measures. It is usually difficult to know how to go about sampling these size-of-effect factors at random to provide broad generalizations. The bottom line is that generalizations of causal inferences are relatively infrequently accomplished by random sampling. And, even then, the permissible generalizations are very limited.

Reasonable external validity can often be obtained by conducting multiple studies where treatments, participants, times, settings, and outcome measures vary across the studies. In addition, treatments, participants, times, settings, and outcome measures can vary even in single studies. Assessing how that variation is related to outcomes can provide information to assist with the generalization of results. In any case, efforts should be made to ensure that samples of treatments, participants, times, settings, and outcome measures are drawn from the populations of greatest interest. For example, Amazon's Mechanical Turk (Mason & Suri, 2012) assists researchers in sampling participants from a broader and more interesting population than is traditionally achieved in laboratory studies using college undergraduates.

In addition, Shadish et al. (2002, p. 353) present "five simple principles that scientists use in making generalizations."

1. *Surface similarity*. Researchers make generalizations based on similarities between the results at hand and the circumstances to which they wish to generalize. That is, researchers look for close matches between the size-of-effect factors in a given study and the prototypical features for which conclusions are sought. For example, in a study of religious beliefs, researchers feel more comfortable generalizing from a sample of Christians to a population of Christians than to a population of Muslims.
2. *Ruling out irrelevancies*. Researchers ignore features that are likely irrelevant to the generalization. For example, religion is likely to be relevant for many generalizations but not to inferences about effects on most basic brain functioning.
3. *Making discriminations*. Researchers limit generalizations based on characteristics deemed relevant to different treatment effects. For example, mothers and fathers have been shown to differ in many parenting styles, so researchers are cautious in generalizing across the sexes when studying parenting outcomes.
4. *Interpolation and extrapolation*. Researchers generalize by projecting trends in the obtained data. For example, researchers interpolate results to fourth graders based on results from third and fifth graders. Similarly, researchers extrapolate results from third and fourth graders to fifth graders. West, Aiken, and Todd (1993) explain, for example, how to conduct parametric studies to assess the response surface of the effects of multiple treatment levels.
5. *Causal explanation*. Researchers generalize based on shared causal mechanisms. For example, that secondhand smoking contributes to lung cancer can be

predicted from the causal linkage between firsthand smoking and lung cancer. Researchers identify the causal mediating forces in a given study and generalize to situations where those forces are also present.

Like induction, these principles of generalization cannot be justified logically. The generalization of results (even when derived from random sampling) is always based on potentially fallible premises (Cook, 2014). The five principles given earlier, however, have been validated by experience and, as a result, provide a reasonable theory to guide practice. When used judiciously, the five principles can lead to credible generalizations. As West and Thoemmes (2010, p. 34) explain, "There is no proof that generalization will be achieved in the new settings, but the use of these principles is expected to substantially enhance its likelihood."

3.7 TRADE-OFFS AMONG TYPES OF VALIDITY

There are often trade-offs among the four types of validity. For example, achieving a high degree of internal validity often means sacrificing a degree of external validity. Different research designs (which will be described in subsequent chapters) tend to have different strengths and weaknesses in terms of different types of validity. For example, randomized experiments (see Chapter 4) tend to be strong in internal and statistical conclusion validity but often weak in external validity. Conversely, the pretest–posttest design (see Chapter 6) and the nonequivalent group design (see Chapter 7) tend to be weak in internal validity but strong in external validity. These trade-offs will be described in further detail in subsequent chapters. The point is that achieving validity is often a balancing act. Different designs and different circumstances often dictate which types of validity are well obtained and which are not.

3.8 A FOCUS ON INTERNAL AND STATISTICAL CONCLUSION VALIDITY

Different researchers place different priorities on the four types of validity (Shadish et al., 2002). Campbell and Stanley (1966, p. 5) consider internal validity to be the "sine qua non" of causal inference: "*Internal validity* is the basic minimum without which any experiment is uninterpretable . . ." (emphasis in original). In contrast, Cronbach (1982) values external validity more than internal validity. According to Cronbach, if the research does not address the questions that need to be answered, it does not much matter if the results are internally valid. Cook and Campbell (1979) argue that applied and basic research tend to have different priorities. While they suggest that internal validity is of paramount importance for both applied and basic researchers, Cook and Campbell (1979) suggest that applied research places greater priority on the construct

validity of the outcome measure than on the construct validity of the cause, whereas basic research has the reverse ordering.

Regardless of the priority different researchers place on different types of validity, this volume emphasizes internal and statistical conclusion validity more than construct and external validity. In this volume, I describe differences among designs in terms of external validity, and I mention construct validity when it is relevant. I focus on internal and statistical conclusion validity because I am concerned with different types of quasi-experimental designs, and differences in internal and statistical conclusion validity are what most distinguish different designs. That is, threats to internal and statistical conclusion validity and their means of control differ significantly across quasi-experimental designs. In contrast, threats to construct and external validity and the means to address them are much the same across designs. For example, the means to take account of placebo effects, evaluation apprehension, and hypothesis guessing (which are threats to construct and external validity) are the same across the different types of designs. In addition, internal validity is a concern unique to the task of assessing effects. Many of the concerns of construct and external validity arise in other types of research such as in survey sampling. I agree with Campbell and Stanley (1966) that internal validity is the "sine qua non" of causal inference. Moreover, statistical conclusion validity is closely aligned with internal validity. Such is not the case with construct and external validity. Such a perspective drives the focus of this volume.

3.9 CONCLUSIONS

If research results are to be credible, researchers must take account of plausible threats to validity. No design will be free from all threats to validity, and different quasi-experimental designs are susceptible to different threats to validity. The researcher's task is to determine which threats to the validity of the results from a specific research design are plausible and to try to minimize those that are most damaging.

3.10 SUGGESTED READING

Chambless, D. L., & Hollon, S. D. (2012). Treatment validity for intervention studies. In H. Cooper (Ed.), *The APA handbook of research methods in psychology* (pp. 529–552). Washington, DC: American Psychological Association.
—Provides an overview of threats to validity and means of their control.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Kruglanski, A. W., & Kroy, M. (1975). Outcome validity in experimental research: A reconceptualization. *Journal of Representative Research in Social Psychology, 7,* 168–178.
—Detail alternatives to the Campbellian perspective on validity.

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentation. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (New Directions for Program Evaluation No. 31, pp. 47–66). San Francisco: Jossey-Bass.
—Reviews several alternative conceptualizations of validity.

Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods, 11,* 1–18.

Reichardt, C. S. (2011b). Criticisms of and an alternative to the Shadish, Cook, and Campbell Validity Typology. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing validity in outcome evaluation: Theory and practice* (New Directions for Evaluation No. 130, pp. 43–53). Hoboken, NJ: Wiley.
—Critique the Campbellian typology of validity and provide an alternative.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
—Provides the definitive explication of the Campbellian typology of validity and describes all 37 threats to validity.

4

Randomized Experiments

Properly implemented randomized experiments serve as the "gold standard"—they typically provide the best, unbiased estimates of the magnitude of the treatment effect.

—West and Thoemmes (2008, p. 419)

When correctly implemented, the randomized controlled experiment is the most powerful design for detecting treatment effects.

—Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson (2007, p. 11)

Randomized assignment designs are a bit like the nectar of the gods; once you've had a taste of the pure stuff it is hard to settle for the flawed alternatives.

—Hollister and Hill (1995, p. 134; cited in Bloom, 2005a, p. 11)

When historians of science look back on the 20th century, they will probably view the use of large-scale field experiments as one of the great achievements of the social sciences. For the first time in history, social scientists used experiments to evaluate the effects of social interventions in areas ranging from education to public health.

—Shadish (2002, p. 3)

Overview

In between-groups randomized experiments, participants are randomly assigned to receive different treatment conditions. The design requires that only posttreatment data be recorded, but the power and precision of the statistical analyses can be greatly increased by collecting pretreatment measures that are either entered as covariates in the statistical model or used for blocking or matching. Complications are introduced if some participants in a randomized experiment fail to comply with the treatment condition to which they were assigned. Complications also arise when some participants fail to provide complete data. The analysis of data in the presence of such complications must rely on additional assumptions, some of which cannot be completely verified.

4.1 INTRODUCTION

In between-groups randomized experiments, participants are assigned to treatment conditions at random (see Chapter 10 for other types of randomized experiments). Random does not mean haphazard. Rather, random assignment means the participants are assigned to treatments based on such genuinely random processes as flipping a coin, rolling a die, or selecting from a computer-generated table of random numbers.

Because participants are assigned to treatment conditions at random, a between-groups randomized experiment is not a quasi-experiment. So why does a book about quasi-experiments have an entire chapter on randomized experiments? There are four reasons. First, randomized experiments provide a benchmark with which to compare quasi-experiments. Randomized experiments are considered by many to be the gold standard for estimating effects. This does not mean randomized experiments are without weaknesses. But it does mean that, in many situations, randomized experiments are the best alternative for estimating effects. Hence, randomized experiments provide an appropriate yardstick to assess the effectiveness of quasi-experiments.

Second, randomized experiments provide the best means to introduce many of the statistical analysis options that are also used in quasi-experiments. As a result, my discussion of randomized experiments provides a necessary background for the discussion of quasi-experiments. The present chapter explains the logic of randomized experiments as a foundation from which to understand the logic of the design and analysis of quasi-experiments.

Third, a research study that starts out as a randomized experiment often ends up as a quasi-experiment. This chapter considers common ways in which randomized experiments can be degraded into quasi-experiments. Such designs are called **broken randomized experiments**. The logic behind these experiments is part of the logic of quasi-experimentation. So in understanding broken randomized experiments, readers are furthering their knowledge of quasi-experiments.

Fourth, knowledge accumulates best when using a diversity of methods that have counterbalancing strengths and weaknesses (Mark & Reichardt, 2004). Randomized experiments may be the gold standard in theory, but they have notable weaknesses in practice (Heckman & Smith, 1995). Correspondingly, quasi-experiments have notable strengths. So randomized experiments combined with quasi-experiments often produce better insights than either randomized experiments or quasi-experiments alone. It serves researchers well to appreciate the relative strengths and weaknesses of randomized experiments when used in combination with quasi-experiments.

Fisher (1932, 1935) is widely regarded as the father of the randomized experiment, though randomized experiments had been considered before Fisher. For example, Bloom (2005a, 2008) dates reference to randomized experiments to the 17th century: "Let us take out of the Hospitals, out of the Camps, or from elsewhere 200, or 500, poor People, that have Fevers, Pleurisies, etc. Let us divide them in halfes, let us cast lots, that one halfe of them may fall to my share, and the other to yours; . . . we shall see how

many Funerals both of us shall have" (John Batista van Helmont, 1662, cited in Bloom, 2005a, p. 11). Bloom (2005a) provides an example of a randomized experiment conducted (and not just hypothesized) in 1885, but Fisher is credited as the first to formalize the underlying logic and statistical analysis of data from randomized experiments and hence generally gets credit as the forebear. Fisher's work began literally in the field because it was born from fieldwork in agriculture and was subsequently assimilated into the behavioral and social sciences. The assimilation was so complete by 1963 that Campbell and Stanley's (1963, 1966) landmark work emphasized quasi-experiments partly in reaction to the hegemony that randomized experiments exercised both in and outside the laboratory in the social and behavioral sciences. While Campbell and Stanley's contributions helped legitimize quasi-experiments, they did not overturn the established pecking order. The widespread sentiment both then and now is that randomized experiments are generally superior to quasi-experiments for estimating treatment effects in the social and behavioral sciences in both the laboratory and field. Quasi-experiments are often thought to be a poor stepchild to be used only when randomized experiments are not feasible, but reality is more nuanced, as will become clear as this book proceeds.

It is hard to say if randomized experiments or quasi-experiments are more widely used overall. Both are commonly used in the field, but use differs by substantive area. For example, randomized experiments are still the mainstay in health-related fields as well as in laboratory research in psychology but are relatively infrequent in education beyond the primary levels of schooling (Cook, 2002, 2008a). Cook (2002, 2008a) examines the reasons most often given for employing quasi-experiments rather than randomized experiments but concludes that perceived barriers are far from insurmountable.

4.2 BETWEEN-GROUPS RANDOMIZED EXPERIMENTS

There are several types of randomized experiments (as explained in Chapter 10). As already noted, in between-groups randomized experiments, different participants are assigned at random to different treatment conditions. For example, children might be assigned at random to reading programs; unemployed workers might be assigned at random to job training programs; or the homeless might be assigned at random to case management services. The participants assigned to the different treatment conditions in a between-groups randomized experiment could be individual persons, or they could be groups of people in classrooms, schools, businesses, hospitals, city blocks, or entire cities. In Chapter 10, I introduce other types of randomized experiments where the units assigned to treatments are times, settings, or outcome measures rather than participants. In the present chapter, however, I consider only between-groups randomized experiments where participants are assigned to treatment conditions at random. For convenience, I sometimes shorten the label of between-groups randomized experiments

to just randomized experiments. In either case, all the randomized experiments in the present chapter fall under the rubric of between-groups randomized experiments.

To avoid possible confusion, it is important to distinguish between **random sampling** and **random assignment.** Random sampling means that participants are a sample drawn at random from some larger population, whereas random assignment means that participants, regardless of whether they had been sampled from a larger population, are assigned to different treatment conditions at random. A randomized experiment does not require that participants be sampled from a larger population at random. All that is required is that participants, however they are sampled, are assigned to treatment conditions at random.

Following Campbellian conventions (Campbell & Stanley, 1966; Cook & Campbell, 1979; and Shadish et al., 2002), I will pictorially represent both randomized and quasi-experiments using X's and O's. An X in a pictorial representation of a design denotes the administration of a treatment, and an O denotes an observation. An O could mean an observation is made on either a single variable or on multiple variables at any given point in time. An observation denotes any measurement whether a paper and pencil test, physiological recording, verbal report, unobtrusive observation, or other empirical assessment. In the diagrams used to represent research designs, time flows from left to right in the positioning of the X's and O's. When observations are made more than one time, subscripts on the O's are used to indicate the time at which the observations are taken.

Using this notation, we can represent the simplest between-groups randomized experiment as follows:

R: X O
R: O

The first row in this depiction indicates that a group of study participants receives a treatment (X), and the behavior of the units is subsequently assessed or measured (O). Because it receives the treatment, this group of participants is called the treatment or experimental group. In addition, the treatment is often called an experimental treatment but that need not determine the nature of the treatment—it can be either innovative or standard; the nature of the treatment simply depends on what the researcher wishes to assess. The second row indicates that a second group of study participants does not receive the experimental treatment but is assessed on the same measurement instruments at the same time as were the participants in the treatment condition. This second group of participants might receive an alternative treatment or no special treatment. The absence of an X in the second row of the depiction simply means the group does not receive the experimental treatment. This group of participants is the comparison group. The "R:" in front of each row indicates that the study participants were randomly assigned to the two treatment conditions. This design is called a **posttest-only between-groups randomized experiment.** The effect of the experimental treatment

(compared to the treatment received in the comparison condition) is estimated by comparing the posttest performances (the O's) in the two conditions.

A slightly more elaborate between-groups randomized experiment is also possible:

$$\begin{array}{llll} \text{R:} & O_1 & X & O_2 \\ \text{R:} & O_1 & & O_2 \end{array}$$

The top row in the depiction indicates that a group of study participants is first observed (O_1), then receives a treatment (X), and is observed again (O_2). The bottom row in the depiction indicates that a second group of study participants is observed, does not receive treatment X (but might receive some comparison treatment instead), and is again observed. The "R:" in front of each line again indicates that the study participants are randomly assigned to the two treatment conditions. (It is often best to perform the randomization after the O_1 observation has been collected, so the placement of the "R:" in the notation is not meant to indicate positioning in time.) This design is called a **pretest–posttest between-groups randomized experiment.** The effect of the treatment is assessed by comparing the posttest performances in the two treatment conditions while taking account of the pretest observations, as will be explained shortly.

The observations (the O's) in a pretest–posttest between-groups randomized experiment will sometimes be called pretreatment and posttreatment observations to distinguish the times of measurements. Sometimes, too, for convenience, the O's will simply be called pretests and posttests, although the observations need not be tests in any sense of that word. Or the O's can be called baseline and outcome measures. In addition, multiple pretest and posttest measures could be collected at each time. Thus, writing both "pretest" and "posttest" as singular (rather than plural) is not meant to be restrictive. The pretest and posttest could just as well indicate multiple measures (as in a battery of measures) collected at a given time. That is, the nature of the pretest and posttest depends on the research circumstances and is not circumscribed because of singular names such as "pretest" and "posttest."

Pretreatment and posttreatment observations need not be operationally identical. For example, they need not be the same or parallel forms of a paper-and-pencil test of cognitive abilities. However, as will be explained later, in many (if not most) cases, it will be advantageous to use operationally identical measures at pretest and posttest. In some cases, however, it will not make much sense to use operationally identical measures. For example, in assessing the effectiveness of an educational program to teach calculus, it makes little sense to use an assessment of the ability to perform calculus as the pretest measurement before any instruction in calculus has been given. If the program were aimed at those with no prior knowledge of calculus, all participants would score zero on such a pretest. Instead, it would make more sense to use an assessment of precalculus mathematical ability as the pretest, which would not be operationally identical to a posttest assessment of the ability to perform calculus.

It is important that the pretests either be measured before the treatments are implemented or be traits (such as age or sex) that are uninfluenced by the treatment. Otherwise, taking the pretests into account in the statistical analysis can bias the results, including removing part, or all, of the effect due to the treatment (Rosenbaum, 1984b; Smith, 1957).

The difference between the two between-groups randomized experiments depicted earlier obviously lies in the pretreatment observations. As will be explained shortly, adding a pretreatment observation to the posttest-only between-groups randomized experiment (to create the pretest–posttest between-groups randomized experiment) can (1) increase the precision and power of the analysis, (2) allow for the study of treatment interactions, and (3) help cope with missing data and attrition. Including pretest measures also allows the researcher to see if the treatment groups are initially balanced (which means the treatment groups are similar on pretest measures). One reason for lack of **balance** between the groups in a randomized experiment is that the protocol for assigning groups at random was not followed faithfully (Boruch, McSweeny, & Soderstrom, 1978; Conner, 1977). Another potential reason is that the assignment procedure resulted in an "unhappy" or "unfortunate" randomization whereby the treatment groups differ simply by chance, but the initial random difference on an important pretest observation is substantively large, including large enough to be statistically significant. When unhappy randomization arises, Rubin (2008b) suggests re-randomizing. In either case, including relevant pretest measures as covariates in the analysis model (see Section 4.6) is a recommended strategy for coping with whatever degree of imbalance exists (Senn, 1994).

Because a pretest measure has some advantages, I focus mostly on the pretest–posttest between-groups randomized experiment, rather than the posttest-only between-groups randomized experiment. For simplicity, I also focus on the simple between-groups pretest–posttest randomized experiment with only two treatment conditions and two time points of observation as diagrammed earlier. More complex designs might add significantly to the information provided but are not needed for the purposes of introducing randomized experiments. More elaborate between-groups randomized experiments include adding posttreatment observations at later points in time. This would allow researchers to assess how treatment effects change over time. For example, Murnane and Willett (2011) report a randomized experiment of the effects of career academies on educational and employment outcomes. Posttreatment measures were collected at both a 4-year and an 11-year followup. After 4 years, the programs showed no statistically significant effects on intended outcomes such as academic skills, graduation rates from high school, or enrollment in college. But the 11-year followup showed substantial differences in income between the treatment groups. Other more complex designs include designs with multiple treatment conditions. Designs with multiple treatment conditions include factorial designs (see Kirk, 2009; Maxwell, Delaney, & Kelley, 2018) and designs where different treatment conditions consist of different components of a complex treatment (West, Aiken, & Todd, 1993). The purposes

of studies assessing different components of a complex treatment include determining which component or components of the complex treatment are responsible for its effects (see Shadish et al., 2002, p. 262).

4.3 EXAMPLES OF RANDOMIZED EXPERIMENTS CONDUCTED IN THE FIELD

Boruch et al. (2009) report that in the 1960s fewer than 100 examples of field-randomized experiments were used to assess the effects of social programs in the United States. Since then, however, there has been a renaissance in the use of randomized experiments in fieldwork in the social sciences (Shadish & Cook, 2009). Boruch et al. (2009) estimated that there were 14,000 randomized or possibly randomized experiments in field social research, reported in sources such as the Campbell Collaboration (which was initiated in 2000). Boruch et al. (2009) reported 500,000 randomized experiments in health fields as documented in the Cochrane Collaboration (which was initiated in 1993).

Perhaps a few classic examples will suffice to give a sense of the breadth of topics that have been investigated with randomized experiments in the field. The Student–Teacher Achievement Ratio (STAR) project (also known as the Tennessee Class Size Experiment) randomly assigned students to classrooms containing different numbers of students to assess the effects of class size on mathematics and reading achievement (Finn & Achilles, 1999). The study enrolled thousands of students in hundreds of classrooms. Although the results have been interpreted differently by different researchers, the initial study concluded that small classes led to better academic outcomes. Racial discrimination by employers was investigated by randomly assigning African American-sounding or white-sounding names to fictitious resumes (Bertrand & Mullainathan, 2004). White-sounding names received 50% more callbacks for interviews than African American-sounding names. The Nurses' Health Study suggested the positive effects of hormone replacement therapy on cardiovascular disease in menopausal women based on nonrandomized comparisons. A subsequent randomized experiment, however, demonstrated no positive effect, along with significant negative side effects (Rosenbaum, 2015b). Flay and Collins (2005) describe various randomized experiments of schoolwide campaigns to prevent tobacco, alcohol, and drug use by students. In the Minneapolis Hot Spots Patrol Experiment, extra police surveillance was assigned to parts of the city at random to reduce crime (Sherman & Weisburd, 1995). Lipsey, Cordray, and Berger (1981) reported the results of a randomized experiment to reduce juvenile delinquency by diverting juvenile offenders out of the court system and into social services. Orr, Bloom, Bell, Doolittle, and Lin (1996) used a randomized experiment to investigate the effects of training programs sponsored by the Job Training Partnership Act, which sought to increase employment and earnings of disadvantaged workers. Classic randomized experiments were also mounted to evaluate the effects of

such programs as Head Start and Follow Through, the Manhattan Bail Bond initiative, the National Institute of Mental Health Treatment of Depression Collaborative Research Program, and the New Jersey Negative Income Tax (Bloom, 2005a; Boruch et al., 2009; Shadish & Cook, 2009).

4.4 SELECTION DIFFERENCES

As explained in Section 2.2, the ideal comparison, which is the comparison that defines a treatment effect, is impossible to obtain in practice. In the ideal comparison, nothing varies with the treatments when they are introduced: everything besides the treatments is the same across the treatment conditions. In any comparison that can be obtained in practice, however, something must vary along with the treatments: everything cannot be held the same.

In a between-groups randomized experiment, the participants vary along with the treatments and, hence, are confounded with the treatments. It is easy to see why. Because different participants are assigned to the different treatment conditions, the participants in the two treatment groups are not the same: therefore, they vary with the treatments. Differences in the participants who receive the different treatments are called selection differences. Selection differences are inherent in between-groups randomized experiments because of the very nature of the comparison being drawn. Therefore, selection differences are an ever present confound in between-groups randomized experiments. But the good news is that because participants are assigned to treatment conditions at random, initial selection differences are random, which means initial selection differences do not bias estimates of the treatment effect. That is one of the three primary advantages of randomized experiments compared to quasi-experiments. In between-groups randomized experiments, simple estimates of treatment effects (such as the difference between posttest means in the two treatment conditions) are unbiased by initial selection differences. That is not generally the case in quasi-experiments.

That initial selection differences do not bias estimates of treatment effects does not, however, mean the effects of initial selection differences can be ignored in between-groups randomized experiments. Even though initial selection differences do not bias estimates of the treatment effect, they still alter estimates of treatment effects and so must be taken into account. But because initial selection differences in between-groups randomized experiments are random, their effects can easily be taken into account using simple statistical significance tests and confidence intervals (Little & Rubin, 2000). That is the second primary advantage of randomized experiments compared to quasi-experiments. Without random assignment, as we shall see in later chapters, the statistical procedures required to cope with the effects of selection differences are much more complex and uncertain (and often heavily dependent on unproven assumptions).

The third primary advantage of randomized experiments lies in the precision of treatment effect estimates and the power of statistical significance tests. With random

assignment, the effects of initial selection differences can be reduced as much as desired simply by increasing the sample sizes of participants assigned to the treatment conditions. This means the precision of the estimate of the treatment effect and the power of statistical procedures to detect the presence of treatment effects can be increased as much as desired simply by increasing sample sizes. In contrast, without random assignment, the effects of selection differences might not decrease substantially by increasing sample sizes.

Other confounds, besides selection differences, can be present in between-groups randomized experiments. Later sections in this chapter explain how the noncompliance of participants to treatment condition assignment (also called nonadherence to treatment conditions) and differential attrition of participants from treatment conditions can bias estimates of treatment effects. Differential **history effects** (i.e., different external influences across the treatment conditions besides the treatments—also called a threat to internal validity due to selection by history) and differential **instrumentation effects** (i.e., differences in measuring instruments across the treatment conditions—also called a threat to internal validity due to selection by instrumentation) are also possible. But I will deal with issues one at a time and focus on initial selection differences first—on their own. That is, I will assume, for the moment, perfect compliance to treatment assignments and no attrition of participants from treatment conditions; I will also assume that randomization has been properly implemented, participants respond independently from each other, different participants in the two conditions are treated the same except for receiving different treatment conditions, and so on.

Some writers have said that random assignment to treatment conditions removes other confounds besides initial selection differences, but random assignment of participants to treatment conditions (accompanied by simple statistical procedures) was devised to address only the problems introduced by initial selection differences. Random assignment (accompanied by simple statistical procedures) is only assured of solving the problem of selection differences under assumptions like those given above (such as full compliance to treatment assignment and no differential attrition). Nonetheless, even when all the given assumptions are not met, random assignment can still be advantageous because it will often be easier to credibly take account of selection differences with randomized experiments than with quasi-experiments. When assumptions are violated, the advantages of randomized experiments compared to quasi-experiments are diminished, though not necessarily completely lost.

4.5 ANALYSIS OF DATA FROM THE POSTTEST-ONLY RANDOMIZED EXPERIMENT

Ignoring all sources of difficulties except selection differences, we can analyze the data from the posttest-only between-groups randomized experiment using the following **regression model:**

$$Y_i = \alpha + (\beta_T T_i) + \varepsilon_i \tag{4.1}$$

The model contains two observed variables, Y_i and T_i. The variable Y_i represents the posttest or outcome score for the *i*th participant. (If there is more than a single posttest measure, each could be modeled separately, or they could all be modeled together using multivariate extensions, which go beyond what is necessary at present.) The variable T_i is an indicator variable (also called a dummy variable) representing the treatment assignment for the *i*th participant, where T_i equals 0 if the participant is assigned to the comparison condition and T_i equals 1 if the participant is assigned to the treatment condition. The ε_i term is called the residual or error term and represents the unexplained variance in the outcome scores—that is, the variability in Y_i that is unexplained by the rest of the model. The ε_i variable is not directly observed; it can only be estimated. The model in Equation 4.1 could be fit to data by regressing Y_i onto T_i using **ordinary least squares (OLS) regression.**

Given the above specifications and using OLS estimation, the estimate of α (the Greek letter alpha) in Equation 4.1 is equal to the mean outcome score, Y_i, for the participants in the comparison condition. The estimate of β_T (where β is the Greek letter beta) is the difference in the mean outcome scores, Y_i, between the participants in the treatment and comparison conditions. That is, if $\overline{Y}_T$ is the mean of the outcome scores for the participants in the treatment condition and $\overline{Y}_C$ is the mean of the outcome scores for the participants in the comparison condition, the estimate of β_T is $(\overline{Y}_T - \overline{Y}_C)$. Under the assumptions given above, the estimate of β_T is an unbiased estimate of the average treatment effect. Rather than using regression analysis software, identical results would be obtained by conducting a *t*-test or an **analysis of variance (ANOVA)** that assesses the mean difference in outcome scores between the two treatment groups. I will call Equation 4.1 the ANOVA model.

The precision of the estimate of the treatment effect in the ANOVA model is determined by the within-group variability of the residuals (i.e., by the variance of the ε_i's within the treatment groups, which is denoted σ^2_ε assuming that variance is the same in both treatment groups). Either reducing variability (say, by using more homogeneous samples of participants or more reliable outcome measures) or increasing the sample size increases precision and power (also see Shadish et al., 2002, pp. 46–47, for a list of methods for increasing precision and power). Increasing the sample sizes in the treatment conditions will increase precision and power as much as desired. If the variability among the participants on the posttest scores varies across groups (i.e., if the variance of the residuals differs across the treatment groups), it can be advantageous to have unequal sample sizes in the treatment conditions to maximize precision and power (List, Sadoff, & Wagner, 2010). For example, two-thirds of the participants might be assigned to the treatment condition, and only one-third might be assigned to the comparison condition. Similarly, unequal numbers of participants could be assigned to treatment conditions if the costs of implementing the treatment differed from the costs of implementing the comparison condition (List et al., 2010). For example, if the

cost of implementing the treatment were sufficiently great, it might be beneficial to assign more participants to the comparison condition than to the treatment condition. But there are diminishing returns for increasing the sample size in a single group. The power and precision of the analysis are ultimately limited by the smallest sample size in the two treatment conditions. The same conclusions about sample sizes also hold for the analyses that follow.

4.6 ANALYSIS OF DATA FROM THE PRETEST–POSTTEST RANDOMIZED EXPERIMENT

Adding pretest observations to the design (and thereby converting a posttest-only between-groups randomized experiment into a pretest–posttest between-groups randomized experiment) provides another way to increase the precision and power of the statistical analysis. As already noted, that is one of the great advantages of including pretreatment observations in the design.

For simplicity, I will assume that only a single pretreatment measure is available. But the statistical analysis easily generalizes to multiple pretreatment covariates. I present two ways to include a pretest in the analysis and thereby increase precision and power. The first is the **analysis of covariance (ANCOVA).** The second is **blocking**—also called stratification or subclassification—of which **matching** is a special case. Blocking and matching are described in Section 4.6.4. When the pretreatment and posttreatment variables are operationally the same, other analyses are sometimes possible including change score analysis (see Section 7.3), which give the same results as a repeated measures analysis of variance. But those analysis strategies are not generally as powerful or precise as either ANCOVA or blocking/matching (Reichardt, 1979).

4.6.1 The Basic ANCOVA Model

In the ANCOVA model, the pretreatment score (X_i) is added to Equation 4.1 (the ANOVA model) to produce Equation 4.2 (the ANCOVA model):

$$Y_i = \alpha + (\beta_T T_i) + (\beta_X X_i) + \varepsilon_i \tag{4.2}$$

(It is an unfortunate standard practice to use the letter *X* for two purposes: to represent the experimental treatment in an "O and X" diagram of a research design, and to represent a pretreatment variable. Please do not be confused. Which way *X* is being used will be made clear by the context.)

There are now three observed variables in the model. As in the ANOVA model, Y_i represents scores on the posttest (i.e., the dependent variable). Also, as in the ANOVA model, T_i is an indicator variable representing the treatment assignment for the *i*th participant, where T_i equals 0 if the participant is assigned to the comparison condition

and T_i equals 1 if the participant is assigned to the treatment condition. The third variable, X_i, represents the scores on the pretest. This variable is called a covariate because it is presumed to co-vary with the dependent variable within the treatment groups. As before, the ε_i term is the residual and represents the unexplained variance in the outcome scores—that is, the variability in Y_i that is unexplained by the rest of the model—but note that the ε_i terms in Equations 4.1 and 4.2 are not the same even though they have been given the same label. The error terms differ because some of the variance in the ε_i term in Equation 4.1 is removed by including the X_i variable in Equation 4.2. Estimates of the model parameters can be obtained by regressing Y_i onto both T_i and X_i as specified in Equation 4.2.

In performing this regression, the ANCOVA model fits a regression line in each of the two treatment groups. Equation 4.2 restricts the slopes of the regression lines to be the same in the two groups. That estimated slope is equal to the estimate of β_X. The estimated intercept of the regression line with the *Y*-axis in the comparison group is equal to the estimate of α. The estimate of β_T is the difference in the estimated intercepts between the regression lines in the two treatment groups. As in the ANOVA model, the estimate of β_T is an unbiased estimate of the treatment effect. Again, let $\overline{Y}_T$ be the mean of the outcome scores for the participants in the treatment condition, and let $\overline{Y}_C$ be the mean of the outcome scores for the participants in the comparison condition. Further, let $\overline{X}_T$ be the mean of the pretreatment observations for the participants in the treatment condition, and let $\overline{X}_C$ be the mean of the pretreatment observations for the participants in the comparison condition. Then the estimate of β_T in Equation 4.2 is $(\overline{Y}_T - \overline{Y}_C) - B_X(\overline{X}_T - \overline{X}_C)$, where B_X is the estimate of β_X, $(\overline{Y}_T - \overline{Y}_C)$ is the difference between the posttest means, and $(\overline{X}_T - \overline{X}_C)$ is the difference between the pretest means. Because participants have been assigned to treatment conditions at random, the expected value of $(\overline{X}_T - \overline{X}_C)$ is equal to zero. Hence, the expected value of the treatment effect estimate is simply the expected value of $(\overline{Y}_T - \overline{Y}_C)$, which is the same expected value as in the ANOVA model. That is, in both Equations 4.1 and 4.2, the expected value of the treatment effect estimate is the expected value of $(\overline{Y}_T - \overline{Y}_C)$ (see Section 4.5). (Technically, the probability limits of the two estimates, rather than their expected values, are equal where the probability limits are much the same as expected values in a large sample. But replacing probability limits with expected values produces the same intuition in what follows.) So in the context of a randomized experiment, the two models are attempting to estimate the treatment effect using the same basic quantity—the mean posttest difference between the two groups. The distinction between the two models is that Equation 4.2 makes an adjustment for the initial (random) selection differences between the treatment groups on the pretreatment scores. That is, one of the two treatment groups will start out ahead of the other on the pretest scores (*X*) simply because of random selection differences and including $B_X(\overline{X}_T - \overline{X}_C)$, as part of the treatment effect estimate adjusts for this advantage so that the groups will be more comparable on the posttest scores (*Y*). As will be shown below, this adjustment also serves to increase the precision of the estimate of the treatment effect and the power of statistical significance tests.

The same ANCOVA model will be used in the nonequivalent group quasi-experiment (in Chapter 7), so it is worthwhile to understand the model fully and the best way to do that is pictorially. In turn, to understand ANCOVA, one must first understand regression analysis, and the best way to understand regression analysis is also pictorially. Figure 4.1 presents a scatterplot of data on a pretest and posttest variable in a single group. The pretest scores vary along the horizontal axis, and the posttest scores vary along the vertical axis. Each participant contributes a single score to the plot, and these scores are represented in the figure by small circles. The shape of the scatterplot that results is represented by an ellipse. The question that arises is "When the posttest scores are regressed on the pretest scores, where is the regression line fit in the scatterplot in Figure 4.1? The figure gives two choices (labeled 1 and 2). Many people think the regression line should be the line marked 1, which is the line that goes through the center (i.e., along the major axis) of the ellipse. In fact, the regression line is the line marked 2. Why line 2 is the best choice for the line is explained next.

Regression analysis is a method for predicting posttest scores given the pretest scores. Suppose that you have an infinite population of scores shaped like the ellipse in Figure 4.1 and that you want to predict the posttest score for a person with a pretest score equal to X'. Given a person with a pretest score equal to X', you should predict that person's score on the posttest to be the mean of the posttest scores for all people in the population who have a pretest score equal to X'. That is, the best way to predict the posttest score for a person with a pretest score equal to X' is to find all the people who had pretest scores equal to X', find the mean posttest scores for all those people, and use that mean as the predicted posttest score. The mean of posttest scores for everyone who has a pretest score equal to X' is called the conditional mean of the posttest given X'. For example, suppose you want to predict the posttest score for someone who has a pretest score one standard deviation above average. In this case, you would find all the

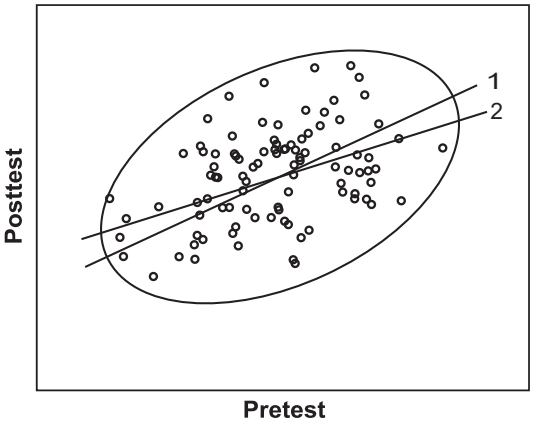


FIGURE 4.1. Two possible lines for the regression of the posttest scores on the pretest scores in a scatterplot.

people who have a pretest score one standard deviation above average and find the average of their posttest scores. This average is called the conditional mean of the posttest for people with pretest scores one standard deviation above average. This conditional mean is the predicted posttest score.

Figure 4.2 divides a scatterplot of scores (represented by the ellipse) into thin slices where each slice contains a thin segment of pretest scores. The mean of the posttest scores in each slice is the conditional mean of the posttest for the given segment of pretests score and is marked by a plus sign. (You can see that the plus signs mark the conditional means of the posttest scores because there are just as many posttest scores above the plus signs as below them, in each thin segment of pretest scores.) These conditional means, represented by the plus signs, are the predicted posttest scores for the given pretest scores. Thus, the line that provides the best prediction of the posttest scores for the given pretest scores runs through these plus signs. The line that runs through the plus signs is drawn in Figure 4.2. Upon inspection, the line in the figure can be seen to be the same as line 2 in Figure 4.1. This is the reason line 2 in Figure 4.1 is the line fit using regression. In other words, you should predict posttest scores, for given pretest scores, using the conditional means of the posttest scores for the given pretest scores. The regression line, which is used to predict posttest scores given pretest scores, runs through these conditional means, as represented by the plus signs and is line 2 in Figure 4.1.

As already noted, when applied to the data from a randomized experiment, an ANCOVA fits a regression line (line 2 in Figure 4.1) to the scores in each treatment group where the posttest scores are regressed onto (i.e., predicted by) the pretest scores. Figure 4.3 presents stylized data from a randomized experiment where the scatters of points from the two treatment groups are represented by the two ellipses. The two regression lines that the ANCOVA in Equation 4.2 fits to the data from the two treatment

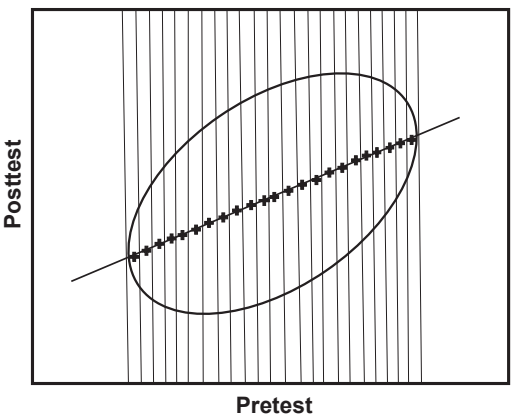


FIGURE 4.2. The regression line drawn through the means of the posttest scores for thin slices of the pretest scores in an idealized scatterplot (indicated by the ellipse). The plus signs are the conditional means of the posttest scores for the slices of pretest scores.

conditions are drawn in Figure 4.3. The estimate of the treatment effect (the estimate of β_T) in the ANCOVA model is equal to the difference in the intercepts of the two regression lines (where the intercepts are the points at which the regression lines meet the *Y*-axis). That difference in the intercepts is also equal to the vertical displacement of one regression line compared to the other. It makes sense to estimate the treatment effect as the vertical displacement between the regression lines. If the treatment has an effect (according to Equation 4.2), it will either raise or lower the posttest scores in the treatment group compared to the posttest scores in the comparison group. Raising or lowering the posttest scores in the treatment group means the regression line in the treatment group would be raised or lowered compared to the regression line in the comparison group. Figure 4.3 depicts a case in which the treatment has a positive effect so that the scatterplot (and the regression line) for the scores from the participants in the treatment condition are raised compared to the scatterplot (and the regression line) from the participants in the comparison condition. The treatment effect would be negative if the regression line in the treatment group was lowered compared to the regression line in the comparison group. That the scatterplot for the treatment condition is not displaced horizontally compared to the scatterplot for the comparison condition is a consequence of random assignment to the treatment conditions. Because of random assignment, the two groups have similar (though, because of random selection differences, not identical)

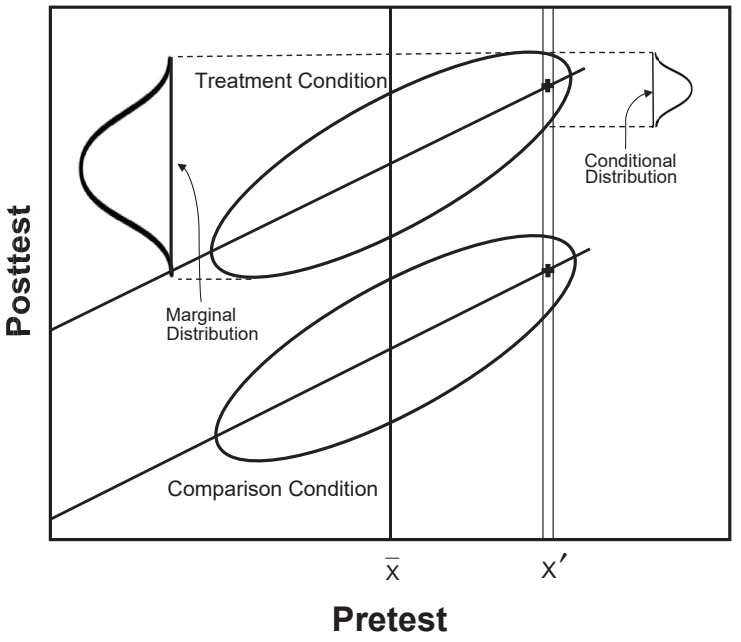


FIGURE 4.3. Idealized scatterplots for a randomized experiment revealing, for each treatment group, the regression lines along with (1) the conditional distribution of the posttest scores given X' in the treatment condition and (2) the marginal distribution of the posttest scores in the treatment condition.

distributions of pretest scores: thus, the groups have scatterplots that are not substantially shifted, compared to one another, along the horizontal axis. The overall mean on the pretest (X) is denoted by $\overline{X}$ in the figure. That the two groups are randomly assigned indicates that the pretest means in the two groups are equal to the overall pretest mean, except for random selection differences.

A thin slice of the data that correspond to the pretest score equal to X' has also been drawn in Figure 4.3. The conditional means of the posttest scores in the two treatment groups for this given X' score are represented in Figure 4.3 by plus signs. As shown in Figure 4.2 and explained above, the regression lines go through these plus signs because the regression lines go through the conditional means of the posttest scores for given pretest scores. Also notice that the vertical distance between these two conditional means is equal to the mean difference in the posttest scores you would find if you matched participants on their pretest scores and compared the means on the posttests for the matched participants. The difference between these two conditional means is also equal to the vertical displacement of the two regression lines, which is equal to the treatment effect. As a result, the ANCOVA estimates the treatment effect in the same way that the treatment effect would be estimated by matching participants on their pretest scores and comparing the posttest scores of those matched participants. That is, ANCOVA is the same as a matching procedure. It is just that ANCOVA performs the matching mathematically rather than by physically matching people based on their pretest scores. It is important to remember that ANCOVA is essentially a matching procedure because it explains how ANCOVA takes account of selection differences in quasi-experiments.

Figure 4.3 also provides a pictorial representation of how the ANCOVA in Equation 4.2 produces treatment effect estimates that are more precise (and statistical significance tests that are more powerful) than does the ANOVA in Equation 4.1. The marginal distribution of the posttest scores in the treatment condition is drawn in the figure as a normal distribution. This distribution is the distribution of the posttest scores that would be obtained if the pretest scores were ignored. The variability of the marginal distribution is equal to the residual variability (σ^2_ε) in the ANOVA model in Equation 4.1. It is this residual variability that determines the precision and power of the ANOVA. Also drawn in Figure 4.3 is a normal distribution representing the conditional distribution of the posttest scores for the given X' scores, in the treatment condition. The variability of this conditional distribution is equal to the residual variability (σ^2_ε) in the ANCOVA model in Equation 4.2. It is this residual variability that determines the precision and power of the ANCOVA. Note how the variability of the conditional distribution (and hence σ^2_ε from the ANCOVA model) is smaller than the variability of the marginal distribution (and hence σ^2_ε from the ANOVA model). The degree to which σ^2_ε from the ANCOVA model is smaller than σ^2_ε from the ANOVA model represents the degree to which the precision and power of the ANCOVA are greater than the precision and power of the ANOVA.

In infinite samples, the relationship between σ^2_ε from the ANCOVA model and σ^2_ε from the ANOVA model is

$$\sigma^2_\varepsilon \text{ from ANCOVA} = [1 - (\rho_{XY})^2]\ \sigma^2_\varepsilon \text{ from ANOVA} \tag{4.3}$$

where ρ^2_{XY} is the pooled squared correlation between the pretest and posttest scores within the treatment groups (see Cochran, 1957, and Reichardt, 1979, for adjustments to this equation in practice because of variability in estimating β_X). The higher is the absolute value of this correlation, the greater is the relative precision and power of the ANCOVA compared to the ANOVA. More than one pretest score (i.e., covariate) can be added to the ANCOVA in Equation 4.2. In that case, the relation between σ^2_ε from the ANCOVA model and σ^2_ε from the ANOVA model is the same as given above except that $(\rho_{XY})^2$ is replaced by the ***R* squared** for the within-group regression of the posttreatment scores on all the pretreatment covariates. To increase power and precision the most, researchers should use pretreatment variables that, as a group, best predict the posttreatment scores within the treatment conditions. Usually but not always, that means including, as one of the covariates, a pretreatment measure that is operationally identical to the posttreatment measure.

4.6.2 The Linear Interaction ANCOVA Model

Note that, in Figure 4.3, the effect of the treatment is constant across the values of the pretest variable. That is, the distance the regression line in the treatment group is raised above the regression line in the comparison group is the same at all values of the pretest. This need not be the case. Figure 4.4 depicts idealized scatterplots in which the regression lines in the two treatment groups are not parallel. In this case, the effect of the treatment is not equal at all values of the pretest. In particular, the effect of the treatment increases with the pretest because the vertical distance between the two regression lines becomes larger as the value of the pretest increases. That the treatment effect varies with the pretest is what is meant by an interaction between the treatment and the pretest variable. The differing slopes of the regression lines in the two treatment groups reflect the degree of the interaction between the treatment and the pretest variable. Because the treatment effect increases linearly with the pretest, the result is a linear interaction.

To fit the data in Figure 4.4, Equation 4.2 must be embellished to become Equation 4.4:

$$Y_i = \alpha + (\beta_T\ T_i) + (\beta_X\ X^*_i) + [\beta_{TX}\ (T_i \times X^*_i)] + \varepsilon_i \tag{4.4}$$

Again, Y_i represents the posttest scores, T_i is the indicator variable representing treatment assignment, and X_i represents the pretest scores. The term X^*_i is equal to (X_i minus $\overline{X}$) where $\overline{X}$ is the overall mean of the X_i scores (i.e., the mean of the pretest scores

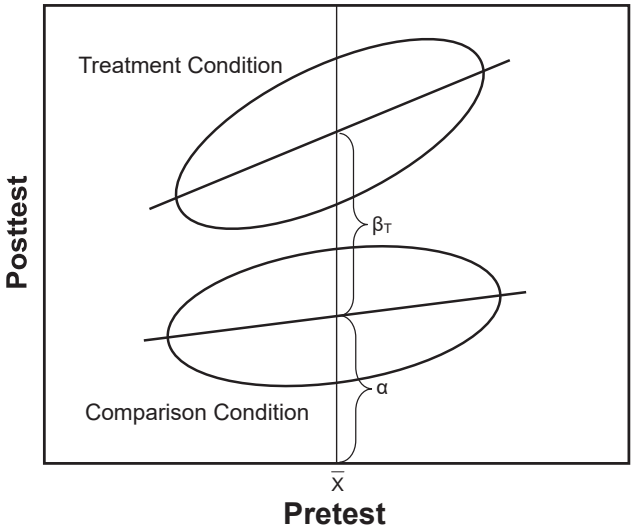


FIGURE 4.4. Idealized scatterplots (with regression lines) for two treatment conditions in the presence of a treatment effect interaction in a randomized experiment.

summed across the participants in both the treatment and comparison conditions). The term $(T_i \times X_i^*)$ is the product of T_i and X_i^*, and ε_i represents the residuals. There are two important differences to note between Equations 4.2 and 4.4. First, X_i in Equation 4.2 has been replaced by X_i^* in Equation 4.4. To be clear, this means that in fitting equation 4.4 to data, $\overline{X}$ is subtracted from X_i and this new variable $(X_i - \overline{X})$ is entered into Equation 4.4 in place of X_i in Equation 4.2. This is called **centering** the pretest scores at the overall mean on the pretest and has the effect of shifting the scaling of the *X*-axis. Figure 4.4 depicts how the data appear. A vertical line is placed in the middle of the scatterplot where *X* equals $\overline{X}$. Subtracting $\overline{X}$ from X_i creates a new *X*-axis where $\overline{X}$ is the zero point. This is equivalent to moving the *Y*-axis to the location of $\overline{X}$. The second change between Equations 4.2 and 4.4 is the addition of the $[\beta_{TX}\,(T_i \times X_i^*)]$ term. With these two changes, the interpretations of the α and β terms also change. The estimate of α is still the estimated intercept of the regression line in the comparison group, but that intercept has been moved to the new *Y*-axis in the middle of the scatterplots (Aiken & West, 1991). That is, the estimate of α is the estimated intercept of the regression line in the comparison group, where the *Y*-axis is placed at the value of $\overline{X}$. The value of α is indicated in Figure 4.4. The estimate of β_T is the estimate of the average treatment effect and is still the vertical displacement between the regression lines. But because the regression lines are no longer parallel, the height displacement is not constant across the pretest (*X*) variable. This means the treatment effect is not constant across the pretest (*X*) variable. By replacing X_i in Equation 4.2 with X_i^* (which is $X_i - \overline{X}$) in Equation 4.4, the treatment effect (β_T) is estimated as the intercept difference between the two regression lines at the new *Y*-axis, which is placed at the value of $\overline{X}$. That is, the treatment

effect (β_T) is the vertical displacement between the regression lines at the point of $\overline{X}$ as depicted in Figure 4.4. As a result, β_T is the average treatment effect (ATE) across all the participants in the sample. That is the reason for substituting $(X_i - \overline{X})$ in Equation 4.4 for X_i in Equation 4.2.

To continue, the estimate of β_X is the estimated slope of the regression line in the comparison group. The estimate of β_{TX} is the estimate of the difference in slopes in the treatment and comparison groups. For example, if the slope of the regression line in the comparison condition is 0.5 and the slope of the regression line in the treatment condition is 1.5, the estimate of β_{TX} is 1. As such, β_{TX} represents the interaction of the treatment with the pretest. A test of the statistical significance of the β_{TX} estimate is a test for the presence of a treatment effect interaction.

You do not have to specify Equation 4.4 using X^*. If you subtract X' from X_i rather than $\overline{X}$ (i.e., if you insert $[X_i - X']$ everywhere into Equation 4.4 in place of $[X_i - \overline{X}]$), the estimate of β_T will be the predicted average treatment effect for the participants with the pretreatment score of X' (Cochran, 1957).

When more than one pretreatment variable is available, these additional pretreatment variables (or covariates) can be added to the ANCOVA model. Further covariates are added by expanding the model in Equation 4.4 to include an additional $(\beta_X X_i^*)$ term for each additional covariate. Interaction terms can also be included for each additional covariate. By adding covariates, the ANCOVA model would mathematically match on all of them all at once. That is, the ANCOVA would estimate the treatment effect by mathematically matching the participants on all the covariates and comparing the posttest scores of those matched participants.

4.6.3 The Quadratic ANCOVA Model

So far, I've assumed that the regressions in the two treatment groups are straight lines. That need not be the case. The ANCOVA model can be fit by adding **polynomial terms** in the pretreatment scores to model nonlinearities in the data. The following is the quadratic ANCOVA model with interaction terms:

$$Y_i = \alpha + (\beta_T T_i) + (\beta_X X_i^*) + [\beta_{TX} (T_i \times X_i^*)] + (\beta_{X2} X_i^{*2}) + [\beta_{TX2} (T_i \times X_i^{*2})] + \varepsilon_i \qquad (4.5)$$

where X_i^{*2} is the square of X_i^* and other variables are defined as in Equation 4.4. In Equation 4.5, the coefficient β_{X2} represents the quadratic relationship between the pretest and posttest scores in the comparison group. The coefficient β_{TX2} represents the difference between the quadratic relationship in the treatment group and the quadratic relationship in the comparison group (which is the quadratic interaction effect). When those two quadratic terms are included in the model, the coefficient β_T is no longer the average effect of the treatment. Instead, β_T is the average treatment effect only for participants who have an X score at the overall mean of the pretreatment scores. (You can estimate the treatment effect for participants with a score of X' by substituting $[X_i$

$- X']$ in place of X_i^*, which is $[X_i - \overline{X}]$, everywhere in the model.) The coefficient β_X is the slope of the tangent to the regression curve in the comparison condition at the value of $\overline{X}$ (Cohen, Cohen, West, & Aiken, 2003). The coefficient β_{TX} is the difference in the tangents to the regression curves in the treatment group compared to the comparison group and so represents the linear interaction of the treatment with X at the value of $\overline{X}$. Adding quadratic terms to the ANCOVA model when the regression surfaces are curvilinear can reduce the error variance and thereby increase the power and precision of estimates of the treatment effect.

Higher-order polynomial terms (such as cubic) could also be added to the model, but there is some uncertainty about whether regression surfaces above quadratic are useful in small to moderate sample sizes. When the regression surface is not well modeled with a quadratic model, procedures such as splines or loess curves are often recommended, though these procedures can also require relatively large sample sizes.

4.6.4 Blocking and Matching

An alternative to ANCOVA is blocking, which is also called stratification or subclassification. In a block-randomized experiment, participants are rank ordered on a pretreatment variable. Then the participants are grouped into blocks or strata based on their rank order on that variable, so similar participants are placed in the same block. For example, if there are to be six participants per block, the six participants with the lowest scores on the pretreatment variable would be assigned to the first block, the six participants with the next lowest scores on the pretreatment variable would be assigned to the next block, and so on. Then the six participants in each block would be randomly assigned to the two treatment groups. As a result of blocking, characteristics on the pretest would tend to be more closely equated across the treatment groups than would be the case with random assignment without blocking.

Any number of blocks can be used. It depends greatly on the number of participants in the study. Five blocks where blocking is based on the quintiles of the pretreatment scores is a popular option (Cochran, 1968b). The goal is to reduce differences on the outcome variable within blocks and maximize differences between blocks.

Equal numbers of participants can be assigned to each treatment condition from within each block or stratum, but this need not be the case. As noted in Section 4.5, an equal number of participants need not be assigned to each treatment condition. Nor do the strata have to have equal numbers of participants.

A special case (called **one-to-one matching**) arises when there is only one participant per treatment condition per stratum. With one-to-one matching, participants are matched based on their pretreatment scores and one participant from each matched pair is assigned randomly to each treatment condition. **One-to-many matching** occurs when each participant assigned to either the treatment or comparison condition is matched to multiple participants who are assigned to the other condition. Randomized experiments with blocking (or matching) are called block-randomized (or matched)

designs as compared to completely randomized designs as in the ANOVA and ANCOVA approaches described earlier which do not involve blocking or matching.

The analysis of data with blocking or matching involves the presence of the blocking or matching in the study. The presence of blocking or matching can be taken into account by adding indicator variables for the blocks or matches to the analysis (Bloom, 2008). If there are K blocks, the analysis could be performed with the following regression model:

$$Y_i = (\beta_T T_i) + (\zeta_1 D_{1i}) + (\zeta_2 D_{2i}) + \ldots + (\zeta_K D_{Ki}) + \varepsilon_i \quad (4.6)$$

where

D_{1i} is an indicator variable coded 1 if the ith participant is in the first block and 0 otherwise;

D_{2i} is an indicator variable coded 1 if the ith participant is in the second block and 0 otherwise;

. . .

D_{Ki} is an indicator variable coded 1 if the ith participant is in the Kth block and 0 otherwise; and

The other variables are as described in previous equations.

Notice that no α (i.e., intercept) term is included in the model (although an intercept could be included using an alternative parameterization where only $K - 1$ rather than K indicator variables were included in the model so one of the block indicators is omitted). The values of ζ_1 to ζ_K (where ζ is the Greek letter zeta) are the effects of the blocks. The size of each ζ term reveals how much participants in that block differ on the outcome variable on average. The other terms in the model (β_T, T_i, and ε_i) are the same as in the ANOVA and ANCOVA models, with the estimate of β_T being the estimate of the average treatment effect. Conceptually, the treatment effect is estimated by calculating mean differences between treatment groups within each block and pooling those mean differences (Bloom, 2008). A researcher can also add terms to assess interactions between the blocks and the treatment, which is like assessing treatment effect interactions in ANCOVA.

Blocking or matching is akin to ANCOVA in that the treatment effect is estimated by comparing participants who are matched on their pretreatment scores. As in ANCOVA, including the blocks or matches in the model can increase the precision and power of the analysis compared to ANOVA. As in ANCOVA, the precision and power of an analysis with blocking or matching increase as the correlation between the pretest and posttest (within the treatment conditions) increases (assuming a linear relationship). Or more specifically, precision and power increase the more similar the posttest scores are within the blocks and the more dissimilar the posttest scores are across the blocks, within each treatment condition. As in ANCOVA, the best pretest measures on which to block

participants are those that correlate most highly with the posttest. The pretest need not be operationally identical to the posttest. But a pretest that is operationally identical to the posttest is often the single measure most highly correlated with the posttest.

Though similar in their goal of increasing power, blocking or matching, and ANCOVA differ in four ways. First, in blocking or matching, the participants in the blocks are only roughly equated on their pretreatment scores. In ANCOVA, the matching is mathematically exact, assuming the regression surfaces have been correctly modeled. For this reason, the increase in precision and power in theory is not quite as great in blocking or matching as in ANCOVA. In both approaches, as noted, precision and power increases as the correlation between the pretreatment variable and the outcome variables increases. If the correlation between the pretreatment variable and the outcome variable is small, however, blocking or matching and ANCOVA could both reduce power because of the loss of degrees of freedom (as explained next). Maxwell, Delaney, and Dill (1984) provide a detailed comparison of the power and precision of blocking or matching compared to ANCOVA.

Second, in blocking or matching, degrees of freedom are lost for each block (i.e., each ζ parameter) included in the model. In parallel fashion, a degree of freedom is lost in the ANCOVA model for each added coefficient (the β parameters). For small sample sizes, the difference in the degrees of freedom between the blocking or matching and ANCOVA could make a difference in precision and power between the two analytical approaches. When simple ANCOVA models fit the data, the differences tend to favor ANCOVA over blocking or matching, though the differences become insubstantial as the sample size increases.

Third, curvilinearity in the regression surfaces could be modeled in ANCOVA using polynomial terms. But if curvilinearity in the regression surfaces is not properly modeled, the standard errors for the treatment effect estimate will tend to be too large. No such model fitting is required in blocking or matching. That is, unlike ANCOVA, nothing needs to be done to deal with curvilinearity in blocking or matching. The blocks or matched pairs automatically conform to curvilinearity in the data (though the more so, the more numerous the blocks). The result is that blocking or matching requires fewer assumptions than ANCOVA.

Fourth, blocking or matching reduces the risks of chance imbalances between the treatment groups (called unhappy or unfortunate randomization) where groups differ substantially on initial characteristics, even though they have been randomly assigned to treatment conditions. Blocking or matching ensures that the treatment groups are reasonably balanced (i.e., are reasonably similar), at least on the covariates that are used for blocking or matching. No such guarantee is provided with ANCOVA and its completely randomized design. Chance imbalances can reduce the face validity of the ANCOVA design and are particularly likely with small sample sizes. Having small samples, however, is also the condition under which loss of power due to loss of degrees of freedom in blocking or matching is most severe.

Like ANCOVA, blocking or matching can be performed with more than a single pretreatment variable. However, blocking and matching becomes correspondingly more complex with additional covariates (Greevy, Silber, & Rosenbaum, 2004). A researcher could block or match on composites of two covariates (e.g., participants could be grouped by sex along with categories of age so as to produce blocks of 25- to 34-year-old females, 25- to 34-year-old males, and so on). Or a researcher could block or match on a composite of the pretreatment variables such as Euclidean or Mahalanobis distance measures (Bloom, 2008). With multiple covariates, however, ANCOVA might well be simpler and easier than blocking or matching. Another option is to combine blocking and ANCOVA (Bloom, 2008; Maxwell, Delaney, & Dill, 1984). The researcher performs blocking or matching and then adds covariates in an ANCOVA.

4.7 NONCOMPLIANCE WITH TREATMENT ASSIGNMENT

For one reason or another, some of the participants randomly assigned to the treatment condition might not receive the treatment. For example, some participants randomly assigned to the treatment condition might refuse treatment, fail to show up for the treatment, or prefer the comparison treatment and seek it out instead of the experimental treatment to which they had been assigned. Such participants are called **no-shows.** Conversely, some of the participants randomly assigned to the comparison condition might end up receiving the treatment condition or services similar to the treatment condition. For example, some participants randomly assigned to the comparison condition might learn of the treatment and either finagle their way into the treatment condition or obtain similar treatment outside the bounds of the research study. Or administrators might violate the random assignment protocol and place some participants into the treatment condition who were originally assigned to the comparison condition. Such participants are called **crossovers.** Both no-shows and crossovers produce what is called noncompliance (or nonadherence) to the treatment assignment. Noncompliance is a threat to internal validity because it can bias the estimates of treatment effects. Noncompliance in a randomized experiment can degrade the design into a broken randomized experiment so that it becomes a quasi-experiment. Therefore, taking account of noncompliance in a randomized experiment can be considered part of the theory of quasi-experimentation.

To maintain the comparability of the participants in the treatment conditions, it is best to try to minimize noncompliance to treatment protocols. Researchers can try to minimize noncompliance by taking such actions as making each treatment condition attractive; removing obstacles (such as transportation difficulties) to participation in the study; providing incentives for participation in the study; and including in the study (before random assignment) only participants who agree to abide by assignment to either treatment conditions (Shadish et al., 2002).

If noncompliance is present, it can be difficult, if not impossible, to estimate the average effect of the treatment across the population of participants in the study (i.e., the average treatment effect [ATE]). Instead, researchers may have to be satisfied with methods that estimate alternative quantities. The present section describes four methods for dealing with noncompliance when it cannot be avoided. In all cases, researchers should document the presence of noncompliance. That is, the researcher should record who did or did not comply in each treatment group. Or more generally, researchers should document the amount and types of treatments received by participants in both treatment groups.

To understand the differences among the four approaches, let Y be the posttest scores. Let T equal 1 if the participant was randomly assigned to the treatment condition and equal 0 if the participant was randomly assigned to the comparison condition. Let D equal 1 if the participants received the treatment (whether or not they were randomly assigned to the treatment condition) and equal 0 if participants did not receive the treatment (whether or not they were randomly assigned to the comparison condition). In the last two approaches to noncompliance, it is assumed that compliance is all or nothing: each participant receives either the full treatment protocol or the full comparison protocol. Alternative approaches are also possible for more complex circumstances (Sagarin, West, Ratnikov, Homan, & Ritchie, 2014), including noncompliance in the presence of missing data (Frangakis & Rubin, 1999).

4.7.1 Treatment-as-Received Analysis

At first glance, it is often appealing, in the presence of noncompliance, to base a comparison on the treatments received rather than on the treatments assigned. That is, a researcher might estimate the effects of the treatment by comparing those who received the treatment to those who did not receive the treatment (i.e., by comparing those with D equal to 1 to those with D equal to 0), rather than by comparing those who were randomly assigned to the different treatment conditions (that is, by comparing those with T equal to 1 to those with T equal to 0). The problem is that such a **treatment-as-received** (also called the "as-treated") approach will tend to be biased, perhaps severely, if those who are no-shows or crossovers are different from those who are not no-shows or crossovers. Consider an example. Perhaps no-shows are those participants who would have performed the worst if they had received the treatment. Also, perhaps crossovers are those who perform the best, among all participants, when they receive the treatment. In that case, a treatment-as-received approach would likely overestimate the effect of the treatment. The opposite pattern of no-shows and crossovers might arise, so the treatment-as-received approach could underestimate the effect of the treatment. And it would typically be difficult, if not impossible, to know which was the case.

As a result of these difficulties, the treatment-as-received analysis is not recommended. But if a treatment-as-received analysis is to be performed, the analysis should be undertaken as if the design were a quasi-experiment (called the nonequivalent group

design, as described in Chapter 7) rather than as a randomized experiment because those who receive the treatment are likely to differ nonrandomly from those who do not receive the treatment. In the nonequivalent group design, the researcher explicitly takes account of selection biases such as those introduced by noncompliance.

4.7.2 Per-Protocol Analysis

Per-protocol analysis is another superficially appealing procedure, but it is not recommended. Per-protocol analysis compares those who completed the treatment protocols as they were originally assigned and discards those who did not complete the treatment protocols as originally assigned. Specifically, per-protocol analysis compares participants who received the treatment ($D = 1$), given that they were assigned to the treatment ($T = 1$), to those who did not receive the treatment ($D = 0$), given that they were not assigned to the treatment ($T = 0$). The problem is this analysis can lead to bias in much the same way that the treatment-as-received analysis can lead to bias. The reason is that the groups being compared are no longer randomly equivalent. It is as if the nonconforming participants had dropped out of the study, so that their data were unavailable (see Section 4.8), where this attrition is not likely to be random. That is, those who fail to complete the treatment condition protocol are likely to be systematically different from those who fail to complete the comparison condition protocol. For example, perhaps those with most need of the treatment drop out of the comparison condition more than from the treatment condition. Or perhaps those who fail to complete the full treatment protocol are those who find the treatment least effective. Or perhaps those experiencing the most negative side effects of treatment in the treatment condition tend to drop out of the treatment more than those in the comparison condition who would experience the most negative side effects were they to have been assigned to the treatment condition. Per-protocol analysis also goes by other names such as on-treatment analysis. By whatever name, the per-protocol estimate of a treatment effect is likely to be sufficiently biased that it is not worth performing.

4.7.3 Intention-to-Treat or Treatment-as-Assigned Analysis

An alternative to the treatment-as-received and per-protocol approaches is to compare participants based on how they were randomly assigned to treatments (the **intention-to-treat (ITT)** or **treatment-as-assigned analysis**) regardless of whether they received the treatment as randomly assigned. That is, the treatment-as-assigned approach compares the $T = 1$ participants to the $T = 0$ participants as the randomized experiment was originally intended to do. The slogan accompanying this method is "analyze them as you've randomized them" (attributed to Fisher, as cited in Boruch, 1997, p. 195).

The resulting estimate of the treatment effect is called the **intention-to-treat (ITT) estimate.** This estimate is the mean difference between the posttest scores in the two groups:

$$\text{ITT estimate} = \overline{Y}_T - \overline{Y}_C \tag{4.7}$$

where $\overline{Y}_T$ is the mean outcome for the treatment group as originally assigned and $\overline{Y}_C$ is the mean outcome for the comparison group as originally assigned. For example (following Bloom, 2008), consider a manpower training program that is intended to increase workers' incomes. If the treatment produced a constant, $1000 increase for everyone, but half the participants were no-shows (with, for simplicity, no crossovers), the average treatment effect which is the ITT effect would be $500. In other words, the treatment raised the participants' income in the treatment group by an average of $500 compared to the outcome in the comparison condition.

As the preceding example suggests, the ITT analysis underestimates the effect of the treatment on those who complied with the treatment assignment (which is the CACE estimate presented below). But the advantage of the ITT analysis compared to the treatment-as-received or per-protocol approaches is that the direction of bias (compared to the complier average causal effect [CAGE] estimate) is known, under the given conditions (see the discussion below). In addition, some researchers argue that the ITT estimate is relevant when a policymaker wishes to know what happens when the treatment is offered and is voluntarily received among the entire population, especially because noncompliance occurs even with presumably mandatory programs. For example, the ITT might help reveal that the treatment is effective among the few people who would accept it if offered, but the treatment might not be accepted by many and so is not very effective overall.

Although the direction of bias in the ITT estimate compared to the CACE estimate is known, the direction of bias compared to the effect with full compliance is unknown. That is, the ITT estimate may either over- or underestimate the effect of the treatment had there been perfect compliance to treatment assignment (Sagarin et al., 2014). For example, if no-shows are those who would benefit most from the treatment, the ITT estimate can underestimate the treatment effect had there been perfect compliance. On the other hand, the ITT estimate can overestimate the treatment effect if those who would have negative effects are noncompliant. The ITT estimate can even have the opposite sign from the full treatment effect—negative when the treatment effect with full compliance is positive and vice versa (Sagarin et al., 2014). In spite of these limitations, the ITT estimate is still widely used and recommended. Indeed, it has been mandated by federal agencies such as the Food and Drug Administration (Ten Have et al., 2008).

The ITT estimate has the advantage of comparing participants who are randomly equivalent. The ITT estimate can be difficult to interpret given that it is the average effect of being offered the treatment compared to not being offered the treatment rather than the average effect of the treatment had everyone received the treatment protocol to which they were assigned. That is, the ITT estimate is the average effect of making the treatment available to those in the treatment group regardless of whether people participated in the treatment (Gennetian, Morris, Bos, & Bloom, 2005, p. 81). Another way

to think of the ITT effect is as the effect of the treatment plus the method of instruction or inducements to take the treatment (Sagarin et al., 2014). With different instructions or inducements, different proportions of the participants might comply with treatment assignment and hence change the ITT estimate. As West and Sagarin (2000, p. 126) note, "Sheiner and Rubin [1995] argue that the ITT estimate of the treatment effect confounds both the efficacy of the treatment with the effectiveness of the instructions to comply with the treatment in the context of the specific experimental trial." If compliance changes (perhaps because the treatment is shown to be effective so that more people are willing to give it a try), the ITT estimate of effect can change. Nonetheless, the ITT estimate should always be reported.

4.7.4 Complier Average Causal Effect

Another approach to addressing noncompliance is to estimate the effect of the treatment just on those who comply with the given treatment protocols whether assigned to the treatment or comparison conditions (Angrist, Imbens, & Rubin, 1996; Bloom, 1984, 2008; Sagarin et al., 2014). This leads to what is called the **complier average causal effect (CACE)** or the **local average treatment effect (LATE).** As presented here, this approach requires the strong assumption that there be no effect of treatment assignment per se. This includes assuming that those who do not receive the treatment when assigned to it would have the same outcome if they had been assigned to the comparison condition. The same holds for those who would cross over to receive the treatment—they would have the same outcome regardless of the treatment condition to which they were assigned. This is called the **exclusion restriction** or assumption (Angrist et al., 1996; Little & Rubin, 2000). In general, for the exclusion restriction to hold, those who are no-shows must receive the same treatment as those in the comparison condition and those who cross over must receive the same treatment as those in the treatment condition. The first part of this restriction is most likely to hold true when the comparison condition consists of no treatment or a standard treatment but less likely to hold true if the comparison condition consists of a nonstandard treatment.

To understand the CACE (or the LATE), we need to distinguish four classes of research participants, which are displayed in Table 4.1. The first class is **compliers.** They accept the treatment to which they are assigned. That is, they receive the treatment if they are assigned to the treatment condition and receive the comparison condition if they are assigned to the comparison condition. (The CACE estimate assesses the effect of the treatment on these participants and these participants alone.) The second class of participants is composed of **always-takers.** They participate in the treatment regardless of the treatment condition to which they were assigned: They are the participants who would be crossovers if assigned to the comparison condition. The third class is composed of **never-takers.** They are the participants who would be no-shows if assigned to the treatment condition: They don't receive the treatment whether assigned to the treatment or the comparison condition. The fourth class is composed of **defiers.**

TABLE 4.1. Different Types of Participants in Treatment and Comparison Conditions

	If assigned to treatment condition	If assigned to comparison condition
Compliers	Treated	Not Treated
Always-takers	Treated	Treated
Never-takers	Not Treated	Not Treated
Defiers	Not Treated	Treated

They refuse whichever treatment they are assigned to: They are both no-shows and crossovers. If they are assigned to the treatment, they are no-shows who participate in the comparison condition. If they are assigned to the comparison condition, they cross over to the treatment condition. The CACE estimate assumes that there are no defiers, which seems plausible in many, if not most, cases. The assumption of no defiers is called the **monotonicity assumption** (Angrist et al., 1996; Little & Rubin, 2000). Note, however, that the CACE estimate does not require the assumption that those who received the treatment are initially equivalent to those who did not. Nor does it require that the treatment effect is a constant for everyone.

First, consider the special case where there are no-shows but no crossovers, so that there are no always-takers (Bloom, 1984). That is, some participants randomly assigned to the treatment ($T = 1$) are no-shows who do not receive the treatment ($D = 0$), but the participants randomly assigned to the comparison condition ($T = 0$) do not cross over to receive the treatment (so for them $D = 0$ as well). In this case, the CACE estimate is

$$\begin{aligned} \text{CACE estimate} &= (\overline{Y}_T - \overline{Y}_C) \,/\, (\text{average } D \mid T = 1) \\ &= \text{ITT estimate} \,/\, (\text{average } D \mid T = 1) \end{aligned} \tag{4.8}$$

where $\overline{Y}_T$ is the mean outcome for the treatment group as originally assigned; $\overline{Y}_C$ is the mean outcome for the comparison group as originally assigned; and (average $D \mid T = 1$) is the proportion of participants randomly assigned to the treatment who receive it. For example (following Bloom, 2008), suppose, for simplicity, that the effect would be a constant $1000 for everyone, but half the participants in the treatment condition were no-shows. Under these conditions, the ITT estimate would be $500, which means the program raised the income of those in the treatment condition by $500 on average. Then the CACE estimate is

$$\begin{aligned} \text{CACE estimate} &= (\overline{Y}_T - \overline{Y}_C) \,/\, (\text{average } D \mid T = 1) \\ &= \$500 \,/\, 0.5 \\ &= \$1000 \end{aligned}$$

Because the treatment was received by only half the participants, the $500 average should be allocated only to that half of the participants. This is what the CACE estimate does under these conditions.

Now consider the general case where there are both no-shows and crossovers and the effect of the treatment is not constant across participants. Note that the distribution of the four groups of participants (in Table 4.1) is the same in the two treatment conditions because of random assignment. Because the always-takers and never-takers contribute equally to the two treatment conditions, they do not affect the ITT estimate. Also note that the average outcome in the treatment condition combines the effects from the compliers and the always-takers. In contrast, the average outcome in the comparison condition due to the treatment effect is due to the always-takers. Subtracting the proportion of always-takers in the comparison condition from the proportion of treatment participants in the treatment condition gives the proportion of compliers. Dividing the ITT estimate by the proportion of compliers gives the average effect of the treatment on the compliers only, which is the CACE estimate:

$$\begin{aligned}\text{CACE estimate} &= (\overline{Y}_T - \overline{Y}_C) \,/\, [(\text{average } D \mid T = 1) - (\text{average } D \mid T = 0)] \\ &= \text{ITT estimate} \,/\, [(\text{average } D \mid T = 1) - (\text{average } D \mid T = 0)]\end{aligned}$$

where $\overline{Y}_T$ is the mean outcome for the treatment group as originally assigned; $\overline{Y}_C$ is the mean outcome for the comparison group as originally assigned; (average $D \mid T = 1$) is the proportion of participants randomly assigned to the treatment who receive it; and (average $D \mid T = 0$) is the proportion of participants randomly assigned to the comparison condition who receive the treatment. So [(average $D \mid T = 1$) – (average $D \mid T = 0$)] is an unbiased estimate of the proportion of compliers.

For example (following Bloom, 2008), suppose the distribution of participants is as given in Table 4.2. (Remember, the CACE estimate assumes that there are no defiers.) Suppose the average outcome among the treated compliers is $1625 and the average outcome among the untreated compliers is $1000, so that the average treatment effect among the compliers is $625. Further suppose that the average outcome among the always-takers (who always receive the treatment) is $*W* (for any value of $*W*), and the average outcome among never-takers (who never receive the treatment) is $WW (for any value of $WW). Then the average outcome in the treatment condition is

$$\overline{Y}_T = (.8 \text{ of } \$1625) + (.15 \text{ of } \$W) + (.05 \text{ of } \$WW)$$

while the average outcome among the comparison condition is

$$\overline{Y}_C = (.8 \text{ of } \$1000) + (.15 \text{ of } \$W) + (.05 \text{ of } \$WW)$$

So the ITT estimate is

$$
\begin{aligned}
\text{ITT estimate} &= (\bar{Y}_T - \bar{Y}_C) \\
&= [(.8 \text{ of } \$1625) + (.15 \text{ of } \$W) + (.05 \text{ of } \$WW)] \\
&\quad - [(.8 \text{ of } \$1000) + (.15 \text{ of } \$W) + (.05 \text{ of } \$WW)] \\
&= .8 \times \$625 = \$500.
\end{aligned}
$$

And the CACE estimate is

$$
\begin{aligned}
\text{CACE estimate} &= (\bar{Y}_T - \bar{Y}_C) \,/\, [(\text{average } D \mid T = 1) - (\text{average } D \mid T = 0)] \\
&= \text{ITT estimate} \,/\, [(\text{average } D \mid T = 1) - (\text{average } D \mid T = 0)] \\
&= \$500 \,/\, (.95 - .15) \\
&= \$500 \,/\, .8 \\
&= \$625
\end{aligned}
$$

which is the average effect that was assumed.

The CACE estimate can also be obtained via an **instrumental variables (IV) analysis** called the Wald (1940) estimator. In the instrumental variable approach, the random assignment (*T*) is the **instrumental variable** (or instrument) for the actual treatment status (*D*). The instrumental variables approach can be implemented using **two-stage least squares (2SLS) regression** (see Section 7.7). In the first stage, the treatment status (*D*) is regressed onto treatment assignment (*T*) to produce a predicted treatment status. In the second stage, the posttest scores (*Y*) are regressed onto the predicted treatment status to produce the estimate of the CACE (which is the regression coefficient for the predicted treatment status). The actual two-stage least squares procedure conducts the two regressions together rather than in separate steps and makes an adjustment for the standard error (compared to the standard error that would be obtained if the procedure were implemented in two separate regression analyses).

The efficiency of the IV approach can be increased by adding covariates to both the equations in the IV model (Gennetian et al., 2005). Also note that the IV procedure is a large sample method: it is biased in small samples. In addition, random assignment must be a "strong" instrument, which means that random assignment must be substantially related to the uptake of the treatment and comparison conditions. That is, random assignment (*T*) must be substantially related to actual treatment status (*D*). The larger

TABLE 4.2. Examples of Percentages of Different Types of Participants in Treatment and Comparison Conditions

	If assigned to treatment condition	If assigned to comparison condition
Compliers	Treated = 80%	Not treated = 80%
Always-takers	Treated = 15%	Treated = 15%
Never-takers	Not treated = 5%	Not treated = 5%

is the percentage of those randomly assigned to the treatment who participate in the treatment and the larger is the percentage of those assigned to the comparison condition who participate in the comparison condition, the stronger is the instrument. Weak instruments produce more biased estimates of the treatment effect. Sagarin et al. (2014) provide references for assessing the CACE under different circumstances, such as when some data are missing, and they discuss alternative ways to estimate treatment effects in the presence of noncompliance.

The drawback to the CACE estimate is that it estimates the average treatment effect only for compliers (that is, only for those who choose to accept whichever treatment is offered). Because the compliers are a subset of all participants, the CACE estimate does not necessarily generalize to all participants. When researchers and stakeholders wish to know what the treatment effect would have been were there full compliance, the CACE estimate will not usually provide that answer unless there was, in fact, full compliance (in which case the ITT estimate will suffice) or unless the treatment had a constant effect on everyone. In addition, the CACE estimate depends on the type of participant who complies under the given study conditions. If the motivation or inducement to comply is altered, the value of the CACE estimate can be altered. Those who comply with treatment assignment in a randomized experiment may differ from those who would comply if a treatment were administered on a larger scale after having been proven on a smaller scale. Nonetheless, the CACE estimate provides a consistent estimate of the treatment effect on a defined subset of participants. If there are no always-takers, the treatment compliers consist of everyone who received the treatment. In that case, the CACE estimate is the average treatment effect on the treated (ATT) or, equivalently, the treatment-on-the-treated (TOT) effect, which can be of greater interest to stakeholders (Angrist & Pischke, 2009, 2015; Bloom, 1984).

4.7.5 Randomized Encouragement Designs

In some research settings, the treatment of interest is available to all qualified candidates, so potential participants in a study cannot be excluded from the treatment at random. For example, administrative mandate might require that a job training program be made available to all those who are out of work. Or a program might have a sufficient budget to provide services to all who are eligible. In such situations of universal availability of the treatment, researchers cannot implement a randomized experiment where the treatment is withheld at random. But it might be possible to encourage (or prompt) some participants, at random, to partake of the treatment and not to encourage (or prompt) others. For example, encouragement might consist of a small incentive, increased knowledge of the availability of the treatment, or improved access to treatment such as by free transportation. In any case, those encouraged to participate in the treatment would presumably do so at a higher rate than those who were not encouraged. A comparison of outcomes of those who were randomly encouraged with the outcomes of those who were not randomly encouraged could then be used to estimate the

effects of the treatment. Such designs are called **randomized encouragement designs** (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2010).

In randomized encouragement designs, encouragement is not likely to be 100% effective. In other words, not everyone who is encouraged to participate in the treatment will do so. Some of those who are not encouraged to participate in the treatment will do so nonetheless. The point here is that there is likely to be a substantial degree of noncompliance with treatment participation as instantiated by encouragement. As a result, the CACE methods in Section 4.7.4 are relevant. Comparison of the outcomes of those who were encouraged with the outcomes of those who were not encouraged produces an estimate akin to an ITT estimate in a standard randomized experiment. For the CACE estimate, the always-takers are those who would partake of the treatment whether encouraged or not; the never-takers are those who would not partake of the treatment whether encouraged or not; and the compliers are those who would partake of the treatment but only if encouraged (also called the **enroll-only-if-encouraged participants**). Then the CACE estimate is an estimate of the average effect of the treatment on the enroll-only-if-encouraged participants. This estimate can clearly vary depending on the nature of the encouragement or prompting. For example, more vigorous encouragement could enlarge the pool of enroll-only-if-encouraged participants and hence alter the size of the average effect (unless the size of the treatment effect is the same for everyone).

The same assumptions required for the CACE estimate in Section 4.7.4 apply to randomized encouragement designs. For one thing, there must be a cohort of participants who are compliers. In the encouragement design, this means encouragement to participate must truly boost participation. In addition, the analysis assumes that encouragement affects the outcomes only via its effect on participation. (In terms of an instrumental variable approach, random assignment to encouragement is the instrumental variable.) This means that participation with encouragement has no greater effect than participation alone. For example, those who are encouraged cannot partake in the treatment more vigorously than the same participant would partake of the treatment if not encouraged. Otherwise, the effect being estimated is the effect of the treatment plus the effect of encouragement rather than the effect of the treatment alone.

4.8 MISSING DATA AND ATTRITION

Missing data causes problems for all designs, including randomized experiments. I will address problems introduced by missing data in the context of randomized experiments, but the results generalize to quasi-experimental comparisons.

Missing data can take several forms. Missing data can arise because participants drop out of a study, which is called attrition. And even if participants do not completely drop out of the study, missing data on outcome variables (because participants fail to provide data on outcome variables for whatever reason) introduces the same problems

as attrition. Alternatively, missing data can arise because some or all participants fail to complete items on a pretest questionnaire or because some or all participants fail to complete either an entire pretest questionnaire or the entire set of pretest measures. The good news with missing data on pretest measures (unlike missing data on posttest measures) is that such missing data do not vitiate the value of random assignment. A researcher can always analyze the data from a randomized experiment without any pretreatment data (see Section 4.5).

In essence, missing data on posttest measures can potentially turn a randomized experiment into a broken randomized experiment, which is a quasi-experiment. So taking account of missing data on posttest measures in a randomized experiment can be considered part of the theory of quasi-experimentation. But note that even in the face of missing data on posttest measures, a randomized experiment is often superior to a nonrandomized study such as a nonequivalent group design (see Chapter 7). Randomization makes the treatment groups similar at least at the start of the study. The groups will stay similar to the extent that missing data is either random in effect, minimal, or nondifferential across treatment groups. In nonrandomized studies, the treatment groups are often very different at the start of a study—which can make the groups differ more than in randomized experiments even after they end up with missing data.

Differential attrition arises when the participants who drop out of the study differ (either in number or kind) across the treatment and comparison groups. Differential attrition can bias the estimate of a treatment effect. For example, if those least (most) likely to benefit from a treatment withdraw from the treatment group more than from the comparison group, the study will no longer be a randomized experiment, and a simple estimate of the treatment effect will likely be positively (negatively) biased.

Because of the potentially biasing effects of attrition, the What Works Clearinghouse (WWC) has established limits on the tolerable degree of attrition that is permitted in a randomized experiment in the field of education (U.S. Department of Education, 2017). The limits depend on the combination of overall and differential attrition. For example, for an overall attrition rate of 5%, differential attrition cannot be more than 6.1% to be tolerable under "cautious assumptions" about the degree of potential bias. To meet WWC design standards, a randomized experiment with a high degree of attrition must either have group differences on baseline measures that are no larger than .05 of the within-group standard deviations or use acceptable statistical adjustments to take account of the effects of baseline differences if those differences are no larger than .25 standard deviations. If group differences on baseline measures are greater than .25 standard deviations, the randomized experiment would not meet WWC design standards.

Attrition can also alter the generalizability of the results of a study. Even if attrition is not differential, attrition can still alter the characteristics of the participants who complete the study compared to those who do not complete the study. For example, even if attrition occurs equally across the treatment groups, if the least cooperative participants drop out of both treatment conditions in the study, the results may not generalize to

anyone beyond the most cooperative participants. That is, when those who drop out differ systematically from those who do not drop out, the final results may not generalize to the original population. And even if attrition is not differential and does not degrade external validity, missing data can still reduce power and precision.

Researchers should assess the overall rate of attrition and the differential rates of attrition across the treatment groups. Comparing the percentage of missing data on outcome variables across the treatment and comparison conditions can provide evidence of a treatment effect, as well as suggest the presence of differential attrition. Assuming there is no missing data on pretest measures, the nature of attrition can be assessed using two-by-two ANOVAs with treatment status as one factor, missing data status on each outcome variable as the other factor, and the pretest scores as dependent variables (Jurs & Glass, 1971). A main effect of treatment versus comparison conditions in this analysis suggests a failure of the randomization procedure. The main effects of missing data status compare the background characteristics of those with missing outcome scores to those without missing outcome scores. Such main effects warn of problems with external validity. Interactions between treatment and missing data status indicate differential attrition across the treatment conditions, which could bias the estimates of treatment effects. (Pretests that are operationally identical to the posttests are particularly useful in speculating about the direction of bias likely due to attrition. For example, if the mean group difference on pretest measures is in the opposite direction as the mean group differences on the outcome measures, attrition does not threaten the estimates of the treatment effects as much as if the two differences are in the same direction.)

Note, however, that the lack of statistically significant results from such analyses does not mean the absence of problems due to attrition because the power of these analyses may be insufficient, and the null hypothesis should never be accepted. Instead, researchers should examine the size of differences and not just their statistical significance. In this way, the analyses should be used to assess the nature and likely degree of attrition—not just its presence or absence. Also note that the nature and degree of attrition is being assessed only on the available pretest measures. The effects of attrition could go undetected because it is present only on unmeasured variables. In addition, West and Sagarin (2000) recommend using **logistic regressions** (to predict attrition status by treatment assignment, by each pretest variable separately, and by the interaction of the two) rather than the ANOVAs in the above procedures, but the logic of the analysis is much the same. Once the nature of attrition has been assessed, it can be dealt with in the following ways.

4.8.1 Three Types of Missing Data

Statisticians have come to distinguish three types of missing data because the three types have different consequences for data analysis (Little & Rubin, 2002). The first type is **missing completely at random (MCAR).** When data are MCAR, it means

missingness on a variable is not related either to the missing values of that variable or to any other variables in the analysis. The latter specification means that no variables in the analysis can predict whether a data point is missing. In essence, MCAR missingness is caused by processes that are random for the purposes of the analysis. If missingness were determined by a random number generator, for example, data would be missing completely at random. Missing data that are MCAR introduce no bias into estimates of treatment effects and do not impact generalizability (but could reduce the power and precision of analyses). You can partially test if missing data are MCAR by looking for differences between those with and without missing data (Little, 1988). But even if the data pass this test, it may be unrealistic to assume missing data are MCAR.

The second type of missing data is data **missing at random (MAR).** When data are MAR, missing scores depend on observed variables but not unobserved variables. Thus, missingness is unrelated to the values of the missing scores once observed variables are included appropriately in the statistical model. Missing data that are MAR means that missingness was essentially determined by a random number generator once the observed variables are properly taken into account in the statistical analysis. For example, missing data could be MAR if data on income were missing solely because those who were less educated failed to report their incomes, if education were properly included in the statistical analysis (Sinharay, Stern, & Russell, 2001). The implication is that missing data that are MAR do not bias estimates of treatment effects if the proper statistical adjustments are made, which might include analyses such as ANCOVA where missingness-related variables are entered as covariates. Data that are MCAR are a special case of data that are MAR. For both MCAR and MAR data, missingness is said to be ignorable because estimates of treatment effects are unbiased by missing data when available data are appropriately included in the statistical analyses. But even when estimates of treatment effects are unbiased, missing data can still cause a reduction in power and precision.

The third type of missing data is data that are **missing not at random (MNAR),** which is sometimes called data not missing at random (NMAR). When data are missing not at random, the probability that scores are missing is related to the missing values themselves (were they available), even after including observed variables in the statistical model. For example, consider a study of substance abuse prevention that measures the degree of substance abuse as an outcome variable. Data would be MNAR if participants with high degrees of substance abuse do not want to report their degree of substance abuse and therefore drop out of the study (so they have no data on the outcome variable). Thus, missingness is related to the missing values on the outcome measure, even after controlling for the other variables in the model. In essence, MNAR data are caused systematically by variables that are unobserved. The problem is that missing data that are MNAR can bias estimates of treatment effects even if all available variables are included in the statistical analysis. For MNAR data, missingness is nonignorable. If sufficient variables were available to be included in the statistical analysis, missing data that are MNAR could potentially be converted into missing data that are MAR. So the

trick in design and analysis is to measure baseline variables that are related to missingness. These variables can therefore be included in the data analyses to make the missing data more like MAR data, thereby reducing bias. Available data provide only indirect ways of testing if missing data are MNAR rather than MAR (Potthoff, Tudor, Pieper, & Hasselblad, 2006).

Approaches to data analysis to cope with missing data can be partitioned into three categories: best practices, conditionally acceptable methods, and unacceptable methods. Even the methods described in the best practices section assume missing data are either MCAR or MAR. Thus, even the best practices are not free of bias if missingness is MNAR. It is difficult to be confident that data are either MCAR or MAR rather than MNAR. Theoretically, it is possible to remove the bias due to missingness that is MNAR, but in practice removing bias requires a model of the causes of missingness (Little, 1995; Mazza & Enders, 2014: Schafer & Graham, 2002; Yang & Maxwell, 2014). Creating such a model is as demanding as analyzing data from the nonequivalent group design, which is described in Chapter 7. In essence, data that are MNAR in a randomized experiment means the experiment has degenerated into a broken randomized experiment and should be treated as if it were a quasi-experiment.

4.8.2 Three Best Practices

Multiple imputation (MI), full information maximum likelihood (FIML), and the **estimation-maximization (EM) algorithm** for missing data are all best practices. All three methods remove bias due to missing data if the data are either MCAR or MAR and all necessary variables are included in the statistical analyses.

Multiple imputation involves three steps (Sinharay et al., 2001). The first step is to create multiple copies of the data wherein missing data are imputed using reasonable guesses of the missing values. Reasonable guesses can be generated using regression models applied iteratively, based on nonmissing data, imputed missing data, and random draws. This step can be accomplished using multivariate imputation by chained equations (MICE), or what is also called sequential regression multiple imputation (Azur, Stuart, Frangakis, & Leaf, 2011). The imputation procedure in the first step is repeated m times to create m datasets, each with different imputed values. The second step is to estimate the treatment effect separately in each of the m datasets. All the variables included in the repeated analyses in the second step must have been included in the regression model in the first step. The third step is to aggregate the results from across the multiple datasets to create a final treatment effect estimate and properly represent uncertainty by estimating the standard error of that estimate. **Auxiliary variables** are variables that are not part of the analysis at steps two or three but predict missingness or are correlated (preferably highly correlated) with variables containing missing data. Although not part of steps two or three, auxiliary variables can nonetheless be included at step one to increase power and precision and help convert otherwise MNAR missingness into MAR missingness and thereby reduce bias (Collins, Schafer, & Kam, 2001;

Mazza & Enders, 2014). Sinharay et al. (2001, p. 328) note, "Even if the missingness mechanism is MNAR, MI may give quite reasonable estimates if there are strong covariates" (where covariates means the variables used to predict the missing values).

Unlike MI, full information maximum likelihood (FIML) analysis (also called direct maximum likelihood or raw-data maximum likelihood) takes account of missing data and treatment effect estimation in a single step. FIML is implemented in many computer programs that perform **structural equation modeling (SEM).** Because ANCOVA and multiple regression methods are a special case of SEM, these computer programs can perform ANCOVA with FIML estimation. The researcher simply specifies the statistical model and requests that the SEM computer program perform a FIML analysis to take account of the missing data. FIML specifies a likelihood function for the data (which is a function giving the probability of the data for a set of unknown parameters, including a parameter for the treatment effect given a statistical model). Then the FIML analysis estimates the unknown parameters by choosing the values that maximize the likelihood function for the given data (Mazza & Enders, 2014). Again auxiliary variables can be included to help convert otherwise MNAR missingness into MAR missingness and thereby reduce bias (Graham, 2003; Mazza & Enders, 2014).

The EM algorithm is an iterative process that reads data case by case and imputes missing values using regression analysis of the available data. During the *E* (estimation) step of each iteration, missing data are imputed using predictions based on the available data. During the *M* (maximization) step of each iteration, the estimates of means, variances, and covariances are produced based on the imputed values. The process is repeated where the regression analyses are based at each iteration on the newly estimated means, variances, and covariances. Iterations continue until the process converges. Auxiliary variables can be included in the algorithm by adding them as part of the variance–covariance matrix that is estimated in order to convert otherwise MNAR missingness into MAR missingness and thereby reduce bias. The EM method produces maximum likelihood estimates of means, variances, and covariances, but they cannot be used to estimate treatment effects without further adjustments (such as is obtained by bootstrapping or using approximate sample sizes) to create standard errors for treatment effect estimates (Allison, 2009; Graham, Cumsille, & Shevock, 2012). As a result, the EM algorithm is a useful method for many purposes, but it might not be as useful for estimating treatment effects as MI or FIML.

4.8.3 A Conditionally Acceptable Method

Listwise deletion (also known as casewise deletion or complete-case analysis) is the most common default option in software programs when dealing with missing data. With listwise deletion, a participant's data are included in the model only if the participant has complete data on all variables in the statistical analysis. The listwise deletion method leads to unbiased estimates of regression coefficients if data are MCAR. Otherwise, listwise deletion tends to introduce bias due to the missing data. With MNAR

data, the bias using listwise deletion is often much the same as with the MI, FIML, and EM methods (Graham et al., 2012). Listwise deletion is often said to be a reasonable option when it would result in a loss of less than 5% of the cases (Graham, 2009). Nonetheless, listwise deletion results in a loss of power and precision and does not allow the use of auxiliary variables to reduce bias when bias is present. So MI, FIML, and EM methods are at least as good at removing bias as listwise deletion and will often be much superior to listwise deletion (Graham, 2009).

4.8.4 Unacceptable Methods

Mean substitution, which is a form of **single imputation**, involves replacing the missing values of a variable with the mean of the variable calculated with the available data on that variable. Other forms of single imputation have also been used, including **hot deck methods** (where missing data are replaced by other values in the dataset) and methods where missing data are imputed using regression-based estimates. Among other problems, data derived by single imputation can introduce bias due to the missing data and tend to produce standard errors for treatment effect estimates that are too small (Widaman, 2006). In any case, single imputation has been superseded by multiple imputation.

In **pairwise deletion** (also called pairwise inclusion or available-case analysis), the researcher uses all the nonmissing data that are available. For example, if participant *P* has missing data on variable *A* but not variables *B* and *C*, the researcher would calculate the correlations between variables *A* and *B* and between *A* and *C* across participants omitting data from participant *P* but calculate the correlation between variables *B* and *C* across participants including data from participant *P*. The main problem with pairwise deletion is that it can introduce bias due to missing data (unless the data are MCAR). Another problem is that pairwise deletion sometimes produces variance–covariance matrices (which are often needed to perform analyses) that are not positive definite, which means statistical methods will fail to run. In addition, accurate standard errors for treatment effect estimates can be hard to come by with pairwise deletion (Allison, 2009; Graham, 2009).

Cohen and Cohen (1985) proposed addressing missing data by using indicator variables to specify missing values. Allison (2002) demonstrated that this technique leads to biased estimates of treatment effects.

4.8.5 Conclusions about Missing Data

Under assumptions such as multivariate normality, good methods (MI, FIML, and EM) exist for coping with missing data that are MCAR or MAR. Coping with missing data that are MNAR requires methods that model the cause of missingness, and these models require additional and untestable assumptions (Mazza & Enders, 2014; Schafer &

Graham, 2002). In fact, if the assumptions of analyses that model missingness to cope with MNAR data are incorrect, these methods can produce even more bias than methods that assume data are MAR (Mazza & Enders, 2014). Unfortunately, there is no way to know if missing data are MAR rather than MNAR. When data are MNAR, the MI, FIML, and EM methods are likely to be less biased than the unacceptable methods described above and might not be grossly biased under plausible conditions (Collins et al., 2001; Schafer & Graham, 2002). In the presence of MNAR data, researchers are advised to conduct multiple analyses, including either the MI, FIML, or EM analyses along with analyses that model an MNAR structure of the data. Researchers can have greater confidence in the treatment effect estimates if the results from the multiple analyses agree, to the extent that the multiple analyses cover the range of plausible assumptions.

The best strategy, of course, is to avoid missing data. Some ways to do so, and especially differential attrition, include the following. Make it easy for participants to complete measurements. In long-term studies, obtain addresses and phone numbers not only from participants but also from families and friends who are likely to know the whereabouts of participants. Track participants over time, which includes keeping in touch with them so that they can be located for follow-up measurements. Provide inducements to encourage participants to complete measurement instruments and to remain in less desirable treatment conditions. Collect data even from those who drop out of the treatment protocol. Include in the study only participants who are willing to serve in either treatment condition. Ribisl et al. (1996), Shadish et al. (2002), and Sullivan, Rumptz, Campbell, Eby, and Davidson (1996) go into greater detail about how to prevent, or at least minimize, missing data.

In spite of a researcher's best efforts, some data in field studies often end up missing. For example, Biglan et al. (1991) report attrition ranging from 5 to 66% in studies of substance abuse prevention. In anticipation of at least some degree of attrition and missing data, researchers should collect data on variables that can predict missingness so that these variables, the researcher hopes, can convert otherwise MNAR missingness into MAR missingness (Collins et al., 2001).

Researchers might also consider trying to bracket the potential effects of missing data (West & Sagarin, 2000). Shadish, Hu, Glaser, Knonacki, and Wong (1998) explain how to create brackets when the outcome variables are dichotomous. Another option is to try to locate meaningful subgroups of respondents that show no attrition, which can thereby supply unbiased estimates of treatment effects for those subgroups (Shadish et al., 2002).

4.9 CLUSTER-RANDOMIZED EXPERIMENTS

In the classic **cluster** (also called group, place, or nested) **randomized experiment,** aggregates of participants (rather than individual participants) are assigned

to treatment conditions at random, with data being available from participants as well as from the aggregates (Bloom, 2005b; Boruch & Foley, 2000; Crespi, 2016; Donner & Klar, 2000). For example, schools (rather than students) could be assigned to treatment conditions at random, and data on both schools and students could be collected. In this example, the aggregates or clusters are schools, the participant data come from students, and the data are called clustered, multilevel, hierarchical, or nested. Aggregates could be any groups of participants such as households, classrooms, businesses, hospital wards, communities, and cities. The point is that, while data are available from participants, only aggregates of participants have been assigned to treatments at random. That makes a difference in how the data are to be analyzed.

Bloom (2005b) reports a cluster-randomized experiment conducted in the field in 1927 (Gosnell, 1927), but such experiments did not become commonplace in social program evaluation until the 1980s. Programs to deter smoking in adolescents have been delivered to classrooms at random (Evans et al., 1981). Aiken, West, Woodward, and Reno (1994) randomly assigned mammography screening assessments to women's groups. Blitstein, Murray, Hannan, and Shadish (2005) assessed the effects of healthy lifestyle interventions assigned at random to entire schools. Murray, Moskowitz, and Dent (1996) assessed the effects of a tobacco use prevention program conducted at the level of communities. Other groups that have been assigned at random include housing projects, physician practices, and hospitals, to name a few (Bloom, 2005a; Raudenbush, Martinez, & Spybrook, 2007). For further examples, see Boruch (2005) and Murray (1998).

Quasi-experiments such as the regression discontinuity design can also involve clustered data (Schochet, 2009). I introduce approaches to clustered data in the present chapter on randomized experiments for convenience. The logic for handling clustered data in quasi-experiments is much the same as that for handling clustered data in randomized experiments.

4.9.1 Advantages of Cluster Designs

There are at least five reasons for using a clustered rather than a nonclustered design (Bloom, 2005b; Cook, 2005). First, sometimes it is more practical or convenient to randomly assign treatments to clusters than to individual participants. For example, a school principal or district superintendent might permit classrooms or schools to be randomly assigned to treatments but not individual students within classrooms or schools.

Second, cluster designs might be useful in avoiding problems such as diffusion or imitation of treatments or other instances of noncompliance to treatment assignment (see Sections 2.4 and 4.7). For example, it tends to be more difficult for a student to cross over to receive an experimental treatment in a different classroom or school than

to cross over to receive an experimental treatment offered to individual students within the same classroom or school.

Third, cluster designs might be needed to avoid externalities or spillover effects where the effect of a treatment has unintended consequences outside the bounds of the intervention. For example, a job training program might reduce the number of locally available jobs for those in a comparison condition. So, rather than increasing overall employment, the program might simply shift employment from nonparticipants to participants, leaving the overall level of unemployment unchanged, even though the program appears to be effective. Such externalities might be avoided by assigning treatment conditions to different cities.

Fourth, some programs are by necessity delivered to people in clusters rather than to individuals. For example, group psychotherapy, a communitywide intervention such as a media campaign for smoking cessation, or a schoolwide reform are inherently programs designed for administration at the level of clusters of people. Sometimes social problems are clustered within locales, such as crime within cities, so that resources are most efficiently delivered in clusters. In addition, the cluster might be the unit to which generalizations are desired (Raudenbush, 1997).

Fifth, effects might be greater when an intervention is applied to all the participants in a cluster rather than to just a few participants. For example, an intervention to improve the reading ability of an entire classroom might have synergistic effects that lead to larger effect sizes than the same intervention applied to just some of the students within a classroom.

4.9.2 Hierarchical Analysis of Data from Cluster Designs

When data are available on both clusters and individuals within clusters, the most sophisticated analyses take account of the data at both levels. Such analyses are most often conducted using what are variously called **hierarchical linear models (HLM),** multilevel models, or random coefficient models. The simplest such models have two levels—clusters and participants nested within clusters. More complex models can have more than two levels with, say, students nested in classrooms that are nested in schools that could be nested in school districts (which were assigned to treatments at random). Each level of data has a separate model.

Consider the simple two-level model where, for example, classrooms are randomly assigned to treatment conditions and the researcher has data from students within classrooms (and perhaps additional data are also available at the level of the classroom). In essence, the analysis calculates mean outcomes at the cluster (e.g., classroom) level, along with means of the cluster means within the treatment and comparison conditions. The analysis takes account of the variation of the individual participants (e.g., students) around the cluster (e.g., classroom) means, as well as the variation of the cluster (e.g., classroom) means around the treatment and comparison group means.

In greater detail, the first (or lower) level of the statistical model (e.g., the model at the level of the student) is

Level 1

$$Y_{ij} = \mu_j + \varepsilon_{ij} \tag{4.9}$$

where

- i is a subscript denoting different participants (e.g., students) within a cluster (e.g., classroom);
- j is a subscript denoting different clusters (e.g., classrooms);
- Y_{ij} is the outcome score for the ith participant (e.g., student) within the jth cluster (e.g., classroom);
- μ_j is the mean of the jth cluster (e.g., classroom) on the outcome score; and
- ε_{ij} is the residual variation of the ith participant's (e.g., student's) score around the jth cluster (e.g., classroom) mean.

This Level 1 model says the outcome scores of the individual participants (e.g., students) vary randomly around their aggregate or cluster (e.g., classroom) mean.

The second (or upper) level of the model is

Level 2

$$\mu_j = \alpha + (\beta_T T_j) + r_j \tag{4.10}$$

where

- μ_j is the jth cluster (e.g., classroom) mean on the outcome score estimated from Level 1 of the model;
- T_j is an indicator variable with value 0 for clusters (e.g., classrooms) in the comparison group and 1 for clusters (e.g., classrooms) in the treatment group;
- α is the mean of the comparison cluster (e.g., classroom) means;
- β_T is the difference between the treatment and the comparison group cluster (e.g., classroom) means; and
- r_j is the residual variation of the jth cluster (e.g., classroom) mean around its treatment or comparison group mean.

The Level 2 model allows the clusters (e.g., classrooms) in the treatment condition to have a different outcome mean than the clusters (e.g., classrooms) in the comparison condition. The HLM analysis fits both levels of the model simultaneously. The estimate of β_T is the estimate of the average effect of the treatment on the clusters (e.g., classrooms).

4.9.3 Precision and Power of Cluster Designs

The power and precision of the statistical analyses are related to the **intraclass** (or intracluster) **correlation (ICC),** which is the proportion of total variance in outcome scores that is due to between-cluster variability. That is, the ICC is the ratio of the variance of the r_j terms to the sum of the variances of r_j and ε_{ij} terms. There are two sources of between-cluster variability (Raudenbush, 1997). First, because of nonrandom assignment to clusters, participants with similar traits can end up in the same clusters creating between-cluster differences. Second, shared conditions within the clusters can cause participants to perform similarly. In either case, the higher is the ICC, the lower are power and precision, if everything else (such as sample size) is the same. As a result, a higher ICC implies that a larger sample size is needed to obtain adequate power and precision.

The power and precision of the clustered analysis are more related to the number of clusters than to the number of participants within the clusters (Bloom, 2005b, 2008; Cook, 2005; Raudenbush & Bryk, 2002). For example, given a total of 1000 participants, power would be .75 if there were 50 clusters of 20 participants each, whereas power would be only .45 with 20 clusters of 50 participants each, assuming an ICC of .1, a standardized effect size of .2, and an alpha level of .05 (List, Sadoff, & Wagner, 2010). The implication is that adding a new cluster tends to add to power and precision more than adding the same number of participants to preexisting clusters. The problem, of course, is that clusters tend to be far more expensive to add than participants within clusters. For further discussion of power and precision in clustered designs, see Bloom (2005b) and Raudenbush and Liu (2000).

4.9.4 Blocking and ANCOVA in Cluster Designs

Just as with the nonclustered randomized experiment (see Section 4.6), using ANCOVA or blocking can dramatically increase precision and power in clustered-randomized experiments (Bloom, 2005b; Bloom, Richburg-Hayes, & Black, 2005; Raudenbush et al., 2007; Schochet, 2008). A researcher can add covariates at either level of the HLM analysis, but power and precision tend to be most sensitive to adding covariates at the cluster level, especially when (1) the ICC is large so that a substantial proportion of the variance lies between rather than within groups, and (2) the covariates are highly correlated with the outcome measures. The covariates that tend to be the most highly correlated with outcome scores are the covariates that are operationally identical to the outcome measures (Bloom, 2005b). The effectiveness of blocking or ANCOVA at the cluster level increases as the ICC increases because, with high ICCs, there is relatively more between-cluster variability that can be modeled and thereby controlled using blocking or ANCOVA.

Adding a cluster-level covariate (X_j), where the covariate can interact with treatment assignment, would produce a Level 2 model of

$$\mu_j = \alpha + (\beta_T T_j) + (\beta_X X_j^*) + [\beta_{TX} (T_j \times X_j^*)] + r_j \quad (4.11)$$

In this equation, X_j^* is equal to $(X_j - \overline{X})$ where X_j is the covariate at the cluster level and $\overline{X}$ is the overall mean of X_j. Again, the estimate of β_T is the estimate of the treatment effect across clusters. Other terms are defined as in previous equations.

Alternatively, researchers could block/match clusters before they are randomly assigned to treatment conditions. Blocking/matching or ANCOVA can be particularly effective in clustered designs when there are relatively few clusters because, as noted above, power and precision of the analysis tend to be low when there are few clusters. But if there are few clusters and if the added covariates are not sufficiently highly correlated with the outcome, blocking and ANCOVA can reduce power and precision because of the loss of degrees of freedom in the statistical analysis. Because blocking reduces the degrees of freedom more than does ANCOVA, blocking might reduce power and precision more than ANCOVA when there are few clusters.

4.9.5 Nonhierarchical Analysis of Data from Cluster Designs

It is possible to have data from clusters without having individual data available from participants within the clusters. It is also possible to have data from participants and to aggregate the data from participants to the level of clusters and then to analyze the data only from the clusters (though the preceding multilevel analysis is generally to be preferred). In both cases, the data from the clusters would be analyzed as if the clusters were individual participants using the methods in Sections 4.5 and 4.6, with weights added to take account of different numbers of persons within clusters. For example, if a researcher had outcome data on students within classrooms, the researcher could aggregate the data from students within the classroom to the level of the classroom. Then the researcher could analyze the aggregate scores (e.g., means) from the classrooms using an ANOVA, ANCOVA, or a block/matching analysis. Such analyses can produce valid confidence intervals and statistical significance tests.

What is not recommended, when data are available at the level of the individual participant, is to analyze the data ignoring their clustered structure. To ignore the clustered structure would mean that the researcher analyzes the data as if the participants were individually assigned to treatments at random when in fact the clusters, rather than the individual participants, were assigned at random. Such an approach generally leads to biased confidence intervals and statistical significance tests. The degree of bias depends on the ICC. If the ICC is zero, there is no bias and the researcher can ignore the hierarchical structure in analyzing the data. But bias arises if the ICC is nonzero because the estimated standard errors would be too small so that confidence intervals would be too narrow and statistical significance tests would be too likely to reject the null hypothesis (Raudenbush, 1997). For example, with 100 participants per cluster, if the ICC is only .033 (so that only 3% of the total variation among outcomes is variation among the clusters), the correct standard error is two times larger than the standard

error from the incorrect analysis that ignores the hierarchical structure of the data (Bloom, 2008). The bias in the standard errors due to assessing individuals rather than groups increases with the ICC. This is why Cornfield (1978, p. 101) wrote, "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception" (cited in Raudenbush, 1997).

4.10 OTHER THREATS TO VALIDITY IN RANDOMIZED EXPERIMENTS

I have considered threats to internal validity due to selection differences, noncompliance, and missing data in the context of nonclustered-randomized experiments. Obviously, clustered-randomized experiments can suffer from these same threats. In addition, there are other threats to validity that can apply to both clustered- and nonclustered-randomized experiments (Shadish et al., 2002). The present section considers such threats. (I might also note that the threats discussed in this section apply to quasi-experiments as well but will not always be repeated in subsequent chapters because they are addressed here.)

Other than receiving different treatments, participants are supposed to be treated the same in all other ways. Such might not be the case in practice. It is possible that external events influence the treatment groups differentially. For example, if the classrooms that implemented the intended treatment implemented additional reforms as well, the additional reforms would be confounded with the intended intervention. Or perhaps the experimental treatment contained unspecified components that were not available in the comparison condition. For example, perhaps a treatment was labeled "talk psychotherapy" when in fact a component of the treatment included medications. In that case, the unspecified components would be confounded with the treatment as labeled. Such confounds are said to be due to differential history, local history, or a selection-by-history interaction.

Differential instrumentation (or a selection-by-instrumentation interaction) is also a potential threat to internal validity. Differential instrumentation occurs when a measuring instrument differs across the treatment conditions. As a result, the measuring instrument is confounded with the treatment conditions and might produce a difference in outcomes independent of a treatment effect. Differential instrumentation could arise if different observers are used in the different treatment conditions. Or differential instrumentation could arise if observers are not kept blind to the treatment conditions so that the expectations of the observers about predicted outcomes of the study play a role in the recording of outcomes. May (2012) notes that differential instrumentation could arise if the observers in a comparison condition became bored or inattentive because the participants in the comparison condition failed to produce the same desired outcomes as in the treatment condition. Differential instrumentation could also arise if observations are based on participant self-reports. For example, participants in the treatment condition might become better than participants in the comparison

condition at producing accurate self-assessments. Pitts, West, and Tein (1996) explain how to assess the presence of differential instrumentation when the treatment affects the factor structure of a dependent measure. According to the classification scheme in Shadish et al. (2002), some of these biases would be classified as due to threats to construct validity rather than to internal validity. But the biasing effects are no less severe. In general, both differential history and differential instrumentation are best dealt with by avoidance rather than by after-the-fact correction.

Other threats to construct validity can also arise (see Section 3.3). For example, placebo effects or other expectancy effects in the participants can be present; the intended treatment might not be implemented with adequate fidelity and strength; or the participants might not be as described. For example, that participants were volunteers might not be made clear in the research report, which would threaten construct validity because interpretations of the results by other stakeholders could be incorrect. Also, the measurement instrument might not be reliable or valid. Either floor or ceiling effects could arise in outcome measurements in one of the treatment groups but might go unrecognized, and so on (see Shadish et al., 2002). Such threats to construct validity must be addressed (and hopefully avoided) by careful design of the research study and careful reporting of the study results.

Statistical conclusion validity can also be threatened in a randomized experiment. The preceding sections of this chapter have explained how to increase the precision of treatment effect estimates and the power of statistical significance tests by using either blocking/matching or ANCOVA rather than ANOVA. These procedures operate by reducing error variance. As already mentioned, there are other, nonstatistical, ways to reduce error variance: using participants who perform homogeneously on outcome measures; settings that promote homogeneous performance; reliable measures; treatments that have the same effects on all participants; and so on. Of course, increasing the sample size serves to increase power and precision in randomized experiments. Yet, power and precision can still be inadequate in any given research setting.

Threats to external validity can also be present especially because steps taken to protect other types of validity can serve to threaten external validity. For example, reducing attrition and noncompliance (to increase internal validity) by including in a study only participants who are willing to abide by their random assignment to treatment conditions can limit the generalizability of the study results. Similarly, increasing power and precision (to increase statistical conclusion validity) by reducing homogeneity among participants (so that they perform more homogeneously on outcome measures) can reduce generalizability to a broader range of participants. In addition, external validity and construct validity can be at odds. Threats to construct validity can be avoided by carefully labeling the five size-of-effect factors, but that does not mean the five size-of-effect factors are the ones desired (which is a question of external validity). For example, being careful to correctly label participants as volunteers can solve problems of construct validity (where failing to provide proper labels might lead stakeholders to assume the

participants were not so restricted in scope), but that does not solve the problem that researchers might not want to limit their generalizations to volunteers.

4.11 STRENGTHS AND WEAKNESSES

Both participants and other stakeholders such as administrators will often resist the random assignment of participants to treatment conditions. Resistance can arise because the advantages of random assignment are not always recognized and the desire to be assigned to the treatment of choice will often be strong. Because of such resistance, administrators can reject the option of random assignment to treatments or undermine randomization after it has been implemented. But methodologists have developed strategies for implementing randomized experiments despite resistance (Shadish et al., 2002). One method is to use a wait-list comparison group. In a design that has such a feature, the participants in the comparison condition are induced to take part in the comparison condition by being promised that they will be given the treatment at a later time, should the treatment prove to be beneficial. The inducement can be effective to the extent that participants in the comparison condition would otherwise not receive the treatment at all.

Murnane and Willett (2011) used an alternative solution to resistance to accept random assignment to treatments in their school-based study. When administrators proved reluctant to allow only some schools to receive additional resources in the treatment condition, Murnane and Willett gave extra resources to all schools but randomly assigned whether the schools received extra resources for their third- or fourth-grade classes. As a result, Murnane and Willett (2011) were able to draw randomized comparisons between third-grade classes that did and did not receive the treatment, as well as randomized comparisons between fourth-grade classes that did and did not receive the treatment. The benefits of random assignment are often worth the extra effort that must go into implementing it.

In the absence of noncompliance and missing data, random assignment of participants to treatment conditions, together with simple statistical procedures, solves the threat to the internal validity of selection differences. This is generally a major advantage of randomized experiments compared to quasi-experiments. But alas, noncompliance and missing data often occur in randomized experiments and thereby convert randomized experiments into broken randomized experiments. Nonetheless, there are statistical methods for coping with at least some forms of noncompliance and missing data. In any case, even in the presence of noncompliance and missing data, randomized experiments can still provide more credible results than can quasi-experiments under many conditions (Shadish & Ragsdale, 1996).

Random assignment could exacerbate certain violations of SUTVA, such as those due to compensatory rivalry or resentful demoralization because of the reactive nature

of randomization itself (see Section 2.4; Fetterman, 1982; Lam et al., 1994). Nonetheless, randomized experiments tend to be stronger on internal validity than are quasi-experiments. The opposite tends to be the case with external validity. That is, quasi-experiments are prone to be stronger on external validity than are randomized experiments. For one reason, only some types of settings tend to permit random assignment, hence restricting the generalizability of results to those types of settings. The same applies to people who would agree to participate in a randomized experiment. That is, some participants who might be willing to participate in a quasi-experiment might not be willing to participant in a randomized experiment—which would again limit the generalizability of results from the randomized experiment. Heckman and Smith (1995, p. 99) call this limit to generalizability randomization bias, which "occurs when random assignment causes the type of persons participating in a program to differ from the type that would participate in the program as it normally operates." In addition, the results of randomized experiments conducted on a small scale might not generalize to results if treatments were offered on a large scale. Finally, there can be ethical and practical constraints on implementing randomized experiments. As a result, quasi-experiments are sometimes the only option when estimating treatment effects.

Randomized experiments often have an advantage over quasi-experiments in terms of statistical conclusion validity because randomized experiments usually need fewer participants to have the same precision and power as quasi-experiments. Numerous methods (such as ANCOVA and blocking/matching) are available for increasing the precision of treatment effect estimates and the power of statistical significance tests in randomized experiments as well as in quasi-experiments.

4.12 CONCLUSIONS

Between-groups comparisons (whether randomized experiments or quasi-experiments—see Chapters 7 and 8) must inevitably cope with the effects of initial selection differences between treatment conditions. Random assignment to treatment conditions makes initial selection differences random. Taking account of the effects of random selection differences can be accomplished with relatively simple statistical methods and with far fewer risky assumptions than taking account of nonrandom selection differences. Such is the substantial advantage of randomized experiments compared to between-group quasi-experiments. The problem is that this advantage can be vitiated by noncompliance and missing data. Statistical procedures can be used to cope with the effects of noncompliance and missing data, but these procedures rest on assumptions that can be unreliable and that are not always testable. When randomized experiments are implemented, researchers need to be careful to ensure that the randomization process is properly implemented and that noncompliance, missing data, and other threats to validity are held to a minimum. According to the What Works Clearinghouse (WWC), well-implemented randomized experiments are eligible to "Meet the WWC

Group Design Standards Without Reservation" (U.S. Department of Education, 2017). But poorly implemented randomized experiments will at best receive a rating of "Meets WWC Group Design Standards with Reservations" and might receive a rating of "Does Not Meet WWC Group Design Standards."

4.13 SUGGESTED READING

Bloom, H. S. (Ed.). (2005a). *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage Foundation.

—Provides an overview of randomized experiments, including discussions of their history, the use of instrumental variables and cluster designs, and empirical comparisons of randomized experiments and quasi-experiments.

Bloom, H. S. (2008). The core analytics of randomized experiments for social research. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 115–133). Thousand Oaks, CA: SAGE.

—Provides further details about analyses to cope with noncompliance and clustering and explains how to calculate the power of statistical analyses for data from randomized experiments.

Boruch, R. F., Weisburd, D., Turner, H. M., III, Karpyn, A., & Littell, J. (2009). Randomization controlled trials for evaluation and planning. In L. Bickman & D. J. Rog (Eds.), *The SAGE handbook of applied social research methods* (2nd ed., pp. 147–181). Thousand Oaks, CA: SAGE.

—Explains the nuts and bolts, with numerous examples, of conducting randomized experiments. The topics covered range from ethical issues to the reporting of results.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York: Cambridge University Press.

—Explains everything you ever wanted to know about the technical details of the modern theory of randomized experiments.

5

One-Group Posttest-Only Designs

In the year of our Lord 1432, there arose a grievous quarrel among the brethren over the number of teeth in the mouth of a horse. For thirteen days the disputation raged without ceasing. All the ancient books and chronicles were fetched out, and wonderful and ponderous erudition, such as was never before heard of in this region, was made manifest. At the beginning of the fourteenth day, a youthful friar of goodly bearing asked his learned superiors for permission to add a word, and straightway, to the wonderment of the disputants, whose deep wisdom he sore vexed, he beseeched them to unbend in a manner coarse and unheard-of, and to look in the open mouth of a horse and find answer to their questionings. At this, their dignity being grievously hurt, they waxed exceedingly wroth; and joining in a mighty uproar, they flew upon him and smote his hip and thigh, and cast him out forthwith. For, said they, surely Satan hath tempted this bold neophyte to declare unholy and unheard-of ways of finding truth contrary to all the teachings of the fathers. After many days of grievous strife the dove of peace sat on the assembly, and they as one man, declaring the problem to be an everlasting mystery because of the grievous dearth of historical and theological evidence thereof, so ordered the same writ down.

—FRANCIS BACON (quoted in Mees, 1934, p. 17; cited in Christianson, 1988, p. 8)

When the social philosopher and one-line comic Henny Youngman was asked how his wife was, he always replied, "Compared to what?"

—SPRINTHALL (1997, p. 179)

Overview

In a one-group posttest-only design, no explicit comparison is drawn between what happens when a treatment is introduced and what happens when a comparison condition is introduced. As a result, there is no explicit way to estimate the effect of a treatment.

5.1 INTRODUCTION

The preceding chapter introduced the randomized experiment, which is often considered the gold standard for estimating effects. The present chapter reverts to one of the least acceptable methods for estimating effects. It is essentially nonempirical, much like the earliest methods for determining the number of teeth in a horse.

I have defined a treatment effect in terms of a difference between what happens after a treatment is implemented and what would have happened if a comparison condition had been implemented instead but everything else had been the same (see Section 2.1). Because a treatment effect is defined in terms of a difference, a treatment effect must be estimated in practice based on a difference: a difference between what happens when a treatment condition is implemented and what happens when a comparison condition is implemented. As a result, all randomized experiments and quasi-experiments estimate a treatment effect using such a difference.

However, treatment effects are sometimes assessed without drawing an empirical comparison between what happens after a treatment condition and a comparison condition are implemented. A one-group posttest-only design is such a nonempirical assessment. In a one-group posttest-only design, observations are collected after a treatment is implemented, but no empirical comparison condition is implemented to provide a counterfactual contrast. Because of the absence of an empirical counterfactual contrast, a one-group posttest-only design has been labeled a pre-experimental design (Campbell & Stanley, 1966).

A **one-group posttest-only design** can be diagramed schematically as

X O

where, once again, X represents a treatment, O represents an observation, and time flows from left to right. In words, in a one-group posttest-only design, a treatment is introduced to a single participant or a single group of participants, an observation is subsequently made, and that observation (O) is taken as the estimate of the treatment effect. That is, it is assumed (rather than assessed empirically) that, in the absence of a treatment effect, the value of the observation (O) would have been zero or some other small number, with any deviation from zero or that small number being taken as evidence of a treatment effect. Thus, a comparison is being drawn, but it is just not a comparison that is derived from direct empirical observation. The obvious problem, of course, is that the posttest observation might not be zero or small in the absence of a treatment effect. In addition, there is no way to check on the plausibility of the assumption without additional data. Consider several examples.

5.2 EXAMPLES OF ONE-GROUP POSTTEST-ONLY DESIGNS

In the late 1960s, 15 students were recruited for a study of police reactions to the Black Panthers, which was a radical political organization known to condone violence (see Huck & Sandler, 1979). Each student had a good driving record and a car that passed a safety inspection. A Black Panthers bumper sticker was attached to each car, and the students were sent off to drive around the city with instructions to drive safely. After 17 days, the students had accumulated 33 traffic citations. The researchers concluded

that the study provided evidence that police were biased against the Black Panthers organization. That is, the study was said to provide evidence that the bumper sticker caused police to give the drivers traffic citations. You will note that no observation was made of the number of traffic citations that would have been received without a Black Panthers bumper sticker. So, it is unconvincing to conclude that the 33 tickets were due to the bumper stickers because just as many tickets might have been given without them. Perhaps, despite the instructions, the students drove wildly to encourage being stopped by police as a way to confirm the study's hypothesis, or perhaps the police were participating in a mass crackdown on traffic violation for everyone.

In the days when pickpockets were hanged for their crimes, it was sometimes found that pickpockets were picking the pockets of observers on the streets who were watching the public hanging of a person convicted of pick pocketing (Sprinthall, 1997). This finding has been used to support the argument that capital punishment for pick pocketing is not effective. Nonetheless, such a conclusion might not be warranted. It may seem unnecessary to add a comparison with how many pick pockets would be plying their trade if no execution were taking place. Such a comparison is necessary, however, because executions might reduce the number of thefts due to pick pocketing, even if executions do not reduce the number all the way to zero.

When a nationwide program to vaccinate people against the swine flu was begun in October 1976, the first group of people given the vaccine were the elderly. The program administered shots to 24,000 elderly people during the first week. Eight states stopped the program because three of those 24,000 died (Freedman, Pisani, & Purves, 1978). The eight states that suspended the vaccination program were presumably attributing the three deaths to the vaccines. However, perhaps just as many people would have died if vaccines had not been given. That is, perhaps at least 3 out of 24,000 elderly people would have died within a week even if they had not received the vaccine. If a researcher had wanted to know if the flu vaccine increased the risk of death, he or she would have had to compare the death rates of those who received the vaccine to the death rates of those who had not received the vaccine. No such comparison is reported here.

5.3 STRENGTHS AND WEAKNESSES

In each of the preceding examples, conclusions were drawn about the effects of treatments based on results from a one-group posttest-only design. Such designs are the easiest of all designs to implement, but one-group posttest-only designs are also likely to provide the least credible estimates of treatment effects. The one-group posttest-only design does not provide an empirical counterfactual comparison of what would have happened if a treatment had not been implemented. Yet, in each example, such comparative data are needed to justify the conclusions. If provided, the comparative data might well shed substantial doubt on the correctness of the conclusions that were reached. Because one-group posttest-only designs are so dependent on the assumption

that an outcome would be zero (or sufficiently small) in the absence of the treatment, the design should be used only when that assumption is highly credible.

Qualitative case studies are sometimes said to be instances where a one-group posttest-only design is used. For example, a novel educational program is implemented in a school, after which a qualitative researcher collects data seemingly in the form of a one-group posttest-only design. Campbell (1975, 1978), however, has argued convincingly that, while such studies may appear to be simplistic one-group posttest-only designs they are not actually. The extensive data that are collected, which typically include in-depth interviews and in-person observations, can provide empirical evidence of what would have occurred in the absence of the treatment. This makes the designs quasi-experimental rather than pre-experimental. In other words, based on their case studies, qualitative researchers can often use observations to establish an empirical counterfactual of what would have happened if a treatment had not been implemented. As a result, in at least some case studies, researchers can make a credible case that a treatment caused an observed outcome.

Consider, however, true one-group posttest-only designs that contain no empirical comparison data and thereby must assume that the obtained posttest observation would be zero (or sufficiently small) in the absence of a treatment effect. It is possible to imagine instances where this assumption is correct: that is, where an outcome could confidently be known to be zero (or sufficiently small) in the absence of a treatment, even without direct observation. For example, consider a treatment that consists of a visiting professor lecturing on the results of an as-yet-unreported study. In that case, participants' knowledge of the study results would be zero in the absence of the lecture. Any knowledge of the study after the lecture can be safely attributed to the effect of the lecture.

But relatively rare are such instances in which it is credible to believe, without additional data, that an observation would be zero (or sufficiently small) in the absence of a treatment. Because the credibility of the one-group posttest-only design rests on the potentially unlikely assumption of a zero (or sufficiently small) outcome in the absence of a treatment, only with great caution should this pre-experimental design be used to estimate treatment effects. In most cases, it would be wise to turn the study into either a quasi-experiment or a randomized experiment by adding an empirical comparison of what would have happened if the treatment had not been implemented.

5.4 CONCLUSIONS

Estimating the effect of a treatment requires a counterfactual comparison. A researcher can obtain a counterfactual comparison by assumption (rather than by empirical observation), but in many, if not most cases, that assumption will be unconvincing. To estimate effects convincingly, researchers usually need to employ either randomized or quasi-experimental designs, as described in other chapters of this volume.

5.5 SUGGESTED READING

Campbell, D. T. (1975). Degrees of freedom and the case study. *Comparative Political Studies, 8,* 178–193.

—Explains how qualitative research may appear to be pre-experimental when it is in fact quasi-experimental.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Skokie, IL: Rand McNally.

—Discusses the one-group posttest-only design in greater detail.

Scriven, M. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 101–118). Beverly Hills, CA: SAGE.

—Explains how to make the most of qualitative studies that employ one-group posttest-only designs.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

—Also discusses the one-group posttest-only design in greater detail.

6

Pretest–Posttest Designs

The "pretest–posttest" experimental design has never been highly regarded as an experimental technique in the behavioral sciences and for good reasons.
—GLASS, WILLSON, AND GOTTMAN (1975, p. 1)

There is something fascinating about science. One gets such wholesome returns of conjecture out of such a trifling investment of fact.
—MARK TWAIN (1883, *Life on the Mississippi*)

Overview

The pretest–posttest design extends the one-group posttest-only design by adding a pretest before a treatment is introduced. The treatment effect is estimated by comparing the difference between pretest and posttest observations. The estimates of treatment effects from such designs are subject to a variety of threats to internal validity. The design is best used in those instances in which the threats to internal validity are least plausible.

6.1 INTRODUCTION

In a basic pretest–posttest design, a pretest measure is observed, the treatment is introduced, and a posttest measure is observed. Such a design can be diagrammed schematically as

$$O_1 \quad X \quad O_2$$

where O_1 is the pretest observation, X is the treatment, O_2 is the posttest observation, and time moves from left to right so that the subscripts on the O's indicate the passage of time. The observations can be made on a single study unit or on multiple study units. That is, the design could be based on a single score at each point in time which could

arise if data were supplied by a single person or a single aggregate of people (such as a classroom, school, city, or state). For example, a researcher could study how a single person changes from pretest to posttest as assessed on a measure of heart rate, a score on a cognitive test of ability, or an attitude about a social issue. Similarly, a researcher could study how a single community changes from pretest to posttest on an aggregated measure of health care insurance costs, performance on standardized achievement tests, or voting for the candidate from a given political party. Alternatively, the design could be applied so that multiple scores are collected at each point in time that would arise if data were supplied by multiple people or multiple aggregates of people. For example, a researcher could study how multiple people change from pretest to posttest as assessed on a measure of heart rates, scores on a cognitive test of ability, attitudes about a social issue, and so on.

In the pretest–posttest design, the treatment effect is estimated as the difference between the pretest and posttest observations. If the data are from a single person or single aggregate of persons, the treatment effect would be the difference between just two numbers, a single pretest score and a single posttest score. Alternatively, if the data are from multiple persons or multiple aggregates of persons, the treatment effect could be calculated as the difference between the means of the pretest and posttest scores across multiple people or multiple aggregates of people.

For the treatment effect estimate to make sense, the pretest and posttest must be measured using the same or parallel instruments. For example, if the observation at posttest is the number of hours worked on a job, the same assessment must be made at pretest. Similarly, if the outcome observation at posttest is a test of math performance, the same or an equivalent test must be used for the pretest. In addition, the design assumes that the posttest observation would remain at the level of the pretest in the absence of a treatment effect.

6.2 EXAMPLES OF PRETEST–POSTTEST DESIGNS

Politicians often use the pretest–posttest design to boast of their successes: After I was elected, they might say, the economy improved and the unemployment rate dropped compared to conditions before I was elected—implying that his or her term in office or policy was the cause of the difference (Redmond, 2016). For example, Campbell and Ross (1968) recount the results of a pretest–posttest comparison where the effects of a crackdown on speeding in Connecticut were reported by then Governor Abraham Ribicoff. In 1968, the year before the govenor imposed the crackdown, the number of traffic fatalities had been 720; that number dropped to 680 the year after the crackdown. Governor Ribicoff attributed the decline to his new ruling, apparently without considering any other potential causes.

Similarly, the results of pretest–posttest designs are often given when newspapers report the effects of interventions. For example, Congress imposed a 55-mile-per-hour

speed limit across the nation in 1974 to reduce the consumption of gasoline. Then in 1985, Congress allowed states to raise their speed limits from 55 to 65 miles per hour on rural interstate freeways. Later, in 1995, Congress removed all speed limit restrictions, whereupon some states raised their speed limits to 75 miles per hour. The effects of these changes in speed limits were reported in the mass media, for example, by comparing the number of traffic fatalities before the change in speed limits to the number of fatalities after the change, even though the differences in fatalities could have been due to other factors (Galles, 1995).

Other examples abound. For instance, St. Pierre, Ricciuti, and Creps (2000, p. 3) report, "The most common design for local evaluations [of the Even Start program], used in 76% of the cases, was to pretest and posttest Even Start families at the start and end of a school year. No control or comparison families were included in these studies."

6.3 THREATS TO INTERNAL VALIDITY

The pretest–posttest design is an obvious extension of the one-group posttest-only design (see Chapter 5). That is, the pretest–posttest design is the same as the one-group posttest-only design except that it has an added observation taken before the treatment is implemented. The pretest observation provides an empirical counterfactual comparison for estimating the treatment effect. That is, the pretest observation is used to assess what would have happened if the treatment had not been implemented, which is juxtaposed with the posttest observation to assess the effect of the treatment. This is a tremendous advance over the one-group posttest-only design, which does not provide an empirical counterfactual comparison. The problem is that the pretest observation in the pretest–posttest design might not well reveal what would have happened if the treatment had not been implemented. The difference between the pretest and posttest (which is the estimate of the treatment effect) may be due to factors other than the treatment, so the estimate of the treatment effect is biased. Such factors are threats to internal validity.

As described in Chapter 3, a threat to internal validity varies (is confounded) with the treatment conditions and could bias the estimate of the treatment effect. In the pretest–posttest design, a threat to internal validity is something, besides the treatment, that changes from pretest to posttest measurements and that could produce a difference between those observations. Depending on whether its effect is positive or negative, a threat to internal validity could cause the estimate of the treatment effect to be either smaller or larger than the actual effect of the treatment. Consider two examples from the Coalition for Evidence-Based Policy (2003). First, a pretest–posttest assessment of the Even Start program would have reported substantial positive changes in school readiness in children from disadvantaged families but a randomized experiment demonstrated that (compared to a comparison condition) the program had no effect (because both treatment conditions exhibited large positive changes over time). Second,

a randomized experiment assessing the effects of the Summer Training and Education Program (STEP), which provided instruction during the summer months for disadvantaged teenagers, found the program to be effective on the outcome measure of reading ability. Although reading ability tends to decline over the summer months (because students are out of school), STEP reduced (though did not eliminate) the amount of decline in the treatment group compared to a randomized comparison group. But because the treatment group still exhibited a decline in reading ability over the summer months (though less of a decline than in the comparison group), a pretest–posttest design would have shown the effect of STEP to be negative rather than positive.

A variety of threats to internal validity that could bias treatment effect estimates are often plausible with pretest–posttest designs. Consider the following eight threats.

6.3.1 History (Including Co-Occurring Treatments)

In the pretest–posttest design, history effects are the effects of external events, besides the treatment, that take place between the pretest and posttest. If historical events cause the posttest observations to differ from the pretest observations, history is a threat to internal validity. For example, in assessing the effects of changes in speed limits on traffic fatalities, changes in the price of gasoline could produce history effects. If the price of gasoline went up, people might drive less, causing fewer accidents. Bloom (2005a, p. 10) notes that the dramatic economic changes of the mid-1990s were history events that "made it difficult to estimate the effects of the federal welfare reform legislation passed in 1996." Or imagine a community intervention designed to increase the public's use of seatbelts. If, between the time of the pretest and posttest, a gruesome accident was widely reported in the media where a celebrity was killed in a car crash because he or she was not wearing a seat belt, that historical event might increase seat belt use even in the absence of an effect of the intended intervention. In that case, the mass media's reporting of such an accident would be a threat to internal validity in the form of a historical event. Or a historical event could be an additional treatment that was implemented along with the intended treatment. For example, the treatment of interest could be an innovative teaching method that was implemented at the same time a whole school reform was implemented. In this case, the intended treatment is confounded with the historical event of the whole school reform.

6.3.2 Maturation

Between the time of the pretest and posttest observations in a pretest–posttest design, people (or other observational units) grow older and perhaps wiser. If the time between the pretest and posttest is short, growing older and wiser may have little effect. But even if the time period is relatively short, the passage of time can still cause people to change such as by becoming more tired, frustrated, bored, or forgetful. They can also become hungrier, which might be relevant if the intervention had anything to do with food or nutrition. Such changes fall under the rubric of **maturation,** which is a natural process

that occurs within the participants because of the passage of time. If these types of maturational changes occur and have nothing to do with the treatment, maturation becomes a threat to internal validity in a pretest–posttest design. For example, a pretest–posttest study of an intervention to address behavioral problems in toddlers could be biased by the changes in behavior, such as learning to speak and locomote effectively, which toddlers might exhibit naturally between the time of a pretest and posttest.

The difference between a history effect and a maturation effect lies in the source of the effect. History effects are external to the participant, while maturation effects are internal to the participant.

6.3.3 Testing

The Heisenberg uncertainty principle implies that doing no more than observing quantum phenomena can change the outcomes being studied. The same applies to humans, although on a macro-level rather than a micro-level. That is, the mere act of observing or measuring humans can change their behavior (Rosnow & Rosenthal, 1997). For example, observing people can make them self-conscious, suspicious, or harder working. In addition, the effects of making observations can change over time as people become used to being observed or because they learn from the observation or data collection process. For example, people can improve their test performance simply by taking a similar version of a test because they become testwise through multiple occasions of test taking. Imagine a study of a novel method of teaching arithmetic to elementary school students. Both the pretest and posttest measures might be multiple-choice tests of basic mathematical procedures. By taking the pretest, the students might learn something about how to take multiple-choice tests, and hence their performance might improve on the posttest for that reason alone. Such practice effects due to becoming testwise are said to occur even on advanced tests such as the SATs.

Or consider a study to assess the effects of a physical fitness program where the pretest and posttest consist of measures of physical fitness, such as running a mile. Based on the pretest assessment, study participants might learn that they need to pace themselves when running a mile rather than starting off at a sprint. By putting that knowledge to use, participants might run a mile faster at the time of the posttest than at the time of the pretest even if the physical fitness program had no effect.

Or consider a study of the effect of a documentary film about the cattle food industry using a pretest–posttest design. The pretest and posttest observations might consist of a survey instrument assessing attitudes about eating meat. After taking the pretest, study participants might be inspired to consider their attitudes about eating meat, which perhaps they had not done previously. Merely considering their habits and attitudes about eating meat based on taking the pretest might cause them to change their subsequent attitudes about eating meat, even if the documentary film had no effect.

Such **testing effects** can be a threat to internal validity in the pretest–posttest design because of the inherent difference between the pretest and posttest observations in the number of prior observations, which has nothing to do with the treatment. By

the time of the posttest observation, the participants in the study have been observed on the pretest. In contrast, no prior observation is taken before the pretest in a pretest–posttest design. If the pretest observation changes how people behave on the posttest, a testing effect is present. In other words, a testing effect arises when the pretest observation changes people's subsequent behavior on the posttest—a change that would have occurred even if the treatment had not been introduced.

6.3.4 Instrumentation

A threat to internal validity due to instrumentation arises in a pretest–posttest design when the measurement instrument used to assess the pretest observations differs from that used to assess the posttest observations. When a change in instrumentation occurs, the difference in the measuring instrument from pretest and posttest can bias the estimate of the treatment effect.

A classic example of a change in instrumentation comes from Orlando Wilson's tenure as chief of police in Chicago in 1959 (Glass, Willson, & Gottman, 1975). Reports of crime increased following Wilson's introduction as police chief simply because Wilson implemented changes in how crimes were reported. Paulos (1988, p. 124) provides another example of a shift in instrumentation over time: "Government employment figures jumped significantly in 1983, reflecting nothing more than a decision to count the military among the employed." Marsh (1985) showed how a change in the measurement of legal penalties for crimes is sometimes accompanied by changes in how the crime is defined. And advances in medical treatments are sometimes accompanied by changes in the categorization of diseases. Such changes in instrumentation can introduce biases in pretest–posttest designs when a treatment is introduced at the same time the nature of the measurements is changed.

A threat to interval validity due to instrumentation can bias results in a pretest–posttest design even if the measurement instrument remains constant but is applied differently from pretest to posttest. For example, consider a pretest–posttest design used to study changes in eating behavior due to a weight-loss intervention. The same self-report instrument might be used for both the pretest and posttest assessments, but participants might differ over time in how they report the food they consumed. For example, perhaps what was perceived as a small portion at the time of the pretest was perceived as a large portion at posttest, or vice versa.

The distinction between threats to internal validity due to testing and due to instrumentation is sometimes confusing. One difference between the threats to internal validity of testing and instrumentation is the following. Testing effects change the level of performance exhibited by the participants between the time of the pretest and posttest. That is, testing effects reflect changes in the participants themselves. In contrast, instrumentation effects change the level of the posttest scores compared to the posttest scores, but these changes do not reflect changes in the participants as much as just changes in the measuring instruments.

6.3.5 Selection Differences (Including Attrition)

The composition of the sample being studied in a pretest–posttest design might change over time, leading to a bias in the estimate of the treatment effect. A common form of composition change is attrition, which is also called experimental mortality. Attrition means the loss of study participants as the study progresses. Conversely, augmentation (Mark, Reichardt, & Sanna, 2000) means the addition of study participants as the study progresses. In the context of a pretest–posttest design, having different participants assessed at the pretest than at the posttest is a threat to internal validity because such compositional changes can cause a difference between the pretest and posttest even in the absence of a treatment intervention. The posttest measurement could be either higher or lower than the pretest, depending on the type of participants who dropped out or were added to the study. For example, if the participants least in need of a remedial intervention (because they were performing well without the program) left the program before they were assessed on the posttest, the estimates of the effects of the program could be biased downward. Conversely, consider a study of an intervention for depression. If those whose depression was greatest at the start of the study choose, out of discouragement, to drop out of the program before it is completed, the estimate of the program's effect could be biased upward. College seniors perform better on tests of academic abilities than do college first year students. Part of the difference arises because not all those who enroll in college graduate. And those who drop out of college tend to be less academically advanced than those who graduate. A threat to internal validity due to changes in sample composition (called a threat to internal validity due to selection differences) can be avoided if the sample is restricted so that data come from the same participants on both the pretest and posttest.

6.3.6 Cyclical Changes (Including Seasonality)

Cycles occur regularly throughout the day and night. Many people tend to be more tired after midday meals than before. Some people tend to be alert in the morning and less so at night; others tend to be the reverse. There are also cycles that occur during the week. Among college students, Mondays are associated with different degrees of attention in class than are Fridays, and partying with alcohol tends to happen on some nights more than on others. There are also monthly and yearly cycles. For example, the length of the day changes regularly with the four seasons along with changes in the weather, which causes predictable changes in people's behavior. Such yearly cycles are said to reflect **seasonality** (and the label of seasonality is also often used to reference cycles other than just those that are yearly).

Cyclical changes can be a threat to internal validity if the pretest is observed during one part of a cycle while the posttest is observed during a different part. For example, consider a 6-month-long psychotherapy program to reduce depression. If the program started during the summer and ended during winter, participants might tend

to become more depressed from pretest to posttest (contrary to what the program was attempting to accomplish) because of increases in depression that frequently accompany winter holidays and the shortening of daylight as occurs in the Northern Hemisphere. Or consider children's levels of physical fitness, which might be greater in the summer (because of outdoor play) than in other seasons. If an exercise program were begun at the start of spring and terminated at the end of the summer, the children's fitness might improve from pretest to posttest not because of the program but because of the natural cycle in children's activity levels across the four seasons.

6.3.7 Regression toward the Mean

Treatments are sometimes given to participants who score on the pretest either above or below typical levels. For example, remedial treatments are often given to participants because they have particularly low scores on a pretest. Or conversely, awards of merit (such as scholarships) tend to be given to participants because they have particularly high scores on a pretest. Under these conditions, the scores on the posttest will tend to be closer to what is typical than were the scores on the pretest, and the difference between the pretest and posttest scores can be a biased estimate of a treatment effect as a result. The reason for the bias is **regression toward the mean** (Campbell & Kenny, 1999; Furby, 1973).

Regression toward the mean occurs because of natural variation between the pretest and posttest. Imagine a series of measurements over time. In typical distributions, the most common outcomes are near the mean. But some observations are higher, and some are lower. If a researcher chose a relatively high observation at one point in time, the most likely outcome that would follow is a lower one, simply because there are more possibilities for a lower outcome than for an even higher outcome. The converse would happen if a researcher selected a relatively low outcome. Then the most likely outcome that would follow is a higher one, simply because there are more possibilities for a higher outcome than for an even lower outcome. That is regression toward the mean. Whenever a treatment is implemented precisely because scores are either high or low on a pretest, researchers should be alert to the possibility of regression toward the mean by the time a posttest is measured, which would bias the estimate of a treatment effect in a pretest–posttest design.

For example, (assuming high scores are desirable) if students are selected because they performed poorly on the pretest, the estimate of the treatment effect in a pretest–posttest design will tend to be larger than it should be because regression toward the mean tends to raise the scores on the posttest even in the absence of a treatment effect. For example, a treatment that had no effect could appear to have a positive effect under these conditions. Conversely, if students are selected because they performed well on the pretest, the estimate of the treatment effect in a pretest–posttest design will tend to be smaller than it should be. For example, a treatment that had no effect could appear to have a negative effect under these conditions.

For a real example, consider the Connecticut crackdown on speeding that was described earlier. The purpose of the crackdown was to reduce the number of traffic fatalities, so the pretest and posttest measures were the number of traffic fatalities the year before and the year after the crackdown, respectively. It is likely that Governor Ribicoff was inspired to introduce the crackdown precisely because the number of fatalities was particularly high in the prior year. If so, a researcher should expect regression toward the mean to occur where the number of traffic fatalities would decline after the crackdown. So the crackdown would appear to have been more effective than it really was. Similarly, a researcher should expect regression toward the mean in a pretest–posttest study of the effectiveness of psychotherapy if people self-select into the program because they are particularly depressed at the start. Such a set of circumstances would also likely make the treatment look more effective than it really is.

6.3.8 Chance

The differences between pretest and posttest could simply be due to an accumulation of **chance differences.** We do not expect traffic fatalities to be identical each year because there are innumerable (essentially random) changes from year to year. Similarly, we don't expect most scores (even physiological ones like blood pressure) to remain perfectly steady over time, even though we cannot put our fingers on all the reasons for the changes. The alternative explanation of chance represents a grab bag of essentially random influences that do not rise to the level of specific history events, maturational changes, testing effects, instrumentation changes, and so on. Nonetheless, we need to entertain chance as a potential explanation for differences over time in the pretest–posttest design.

As noted, in some circumstances, the pretest–posttest design is implemented with a group of participants (say, of size *N*), so there are multiple (i.e., *N*) pretest scores matched with the same number (*N*) of posttest scores. In this case, a researcher can perform a statistical test (such as a **matched-pairs *t*-test**) to help assess the likelihood that chance alone could explain the mean difference between pretest and posttest scores. In designs, such as the Connecticut crackdown on speeding, where there is a single pretest and posttest score, the role played by chance cannot be assessed without further data such as from an interrupted time-series design (see Chapter 9).

6.4 DESIGN VARIATIONS

The basic pretest–posttest design can be usefully supplemented by adding another pretest observation, so the design becomes

$$O_1 \quad O_2 \quad X \quad O_3$$

Adding the second pretest measurement allows the pattern of growth to be better modeled over time. Such modeling can help rule out threats to internal validity. For example, on the one hand, if the levels of performance at Time 1 (O_1) and Time 2 (O_2) are the same, effects due to maturation, testing, and regression toward the mean may not be present—for if they were present, a researcher should expect performance differences at Times 1 and 2. On the other hand, if the levels of performances at Times 1 and 2 are not the same, the researcher might cautiously use the difference between Times 1 and 2 to estimate the difference to expect between Times 2 and 3 due to maturation. For example, if the length of time between Time 1 and Time 2 is the same as the length of time between Time 2 and Time 3, and if maturation is assumed to be constant (i.e., linear) across time, the researcher could estimate the effect of the treatment as a difference of differences. That is, the researcher could estimate the treatment effect as the difference between performance at Times O_3 and O_2 minus the difference in performance between Times O_2 and O_1. That is, the treatment effect would be estimated as any change from Times 2 and 3 that is greater than the change from Times 1 and 2 that is the predicted change due to maturation. However, the assumption that maturation is the same over time is not a trivial one. Changes due to maturation are often not expected to be the same over time. So, the double pretest design is not a panacea for taking account of maturational trends. Taking account of the effects of maturation with the double-pretest design requires assumptions that are not testable solely with the data from the design. Another problem is that other threats to validity, such as testing and regression toward the mean, are not likely to have linear effects across observations over time. Adding even more pretest measurements over time can further strengthen the analysis and increase a researcher's confidence in the assumptions that underlie the design. Indeed, adding even more pretest (and perhaps posttest) observations leads to the interrupted time-series design which is described in Chapter 9. The interrupted time-series design can produce results that are notably more credible than the basic pretest–posttest design, as explained in Chapter 9.

Other design variations are also possible. In the classic pretest–posttest design, the same participants are measured at both pretest and posttest. Alternatively, different participants could be measured at the two time points. But just as the pretest and posttest measures must be the same or parallel measures, the participants assessed at the pretest must be either the same as or like the participants assessed at the posttest. Using **cohorts** for the pretest and posttest samples is a common way to ensure that different groups of participants are similar. Cohorts arise, for example, when waves of people cycle through an institution such as a school. To continue the example, the students who are first graders in Year 1 are a cohort to students who are first graders in Year 2 who are cohorts to students who are first graders in Year 3, and so on. Cohorts might also be used in comparing military inductees, business trainees, or family siblings. The premise is that different cohorts are likely to be similar to each other.

The simplest pretest–posttest cohort design is diagrammed as follows:

$$\begin{array}{lll} \text{NR:} & O_1 & \\ \cdots\cdots\cdots\cdots\cdots\cdots & & \\ \text{NR:} & X & O_2 \end{array}$$

The two separate lines in the diagram indicate that there are two separate groups of participants and that the participants measured at Time 1 are different from the participants measured at Time 2. The "NR:" and the dotted line signify that the different groups of participants are cohorts, so the participants were not assigned to treatment conditions at random. For example, if the cohort measured at Time 1 were first graders in the spring of academic Year 1, the cohort measured at Time 2 could be first graders in the spring of academic Year 2, where the treatment was first introduced in the fall of academic Year 2. The cohort design often produces less credible results than does a pretest–posttest design where observations are collected on the same individuals. But the cohort design has its place. For example, if data are collected in a school every spring, the cohort design could be used to compare students before and after a treatment when they were the same age and in the same grade. For example, for a program that begins in the first grade in the fall of academic Year 2, the spring scores on an achievement test for students who were first graders in academic Year 2 could be compared with the spring scores of students who were first graders in academic Year 1. In contrast, a pretest–posttest design using data from the same students would have to compare students when they were different ages. For example, the spring scores of students who were first graders in academic Year 2 would be compared with the scores of the same students when they were younger (such as in the preceding fall or spring). Yet this would likely make maturation a highly plausible threat to internal validity.

Another variation in the basic pretest–posttest design entails the addition of a nonequivalent dependent variable (Coryn & Hobson, 2011; Shadish et al., 2002), which is something Rosenbaum (2017) calls a control outcome. The design is diagrammed:

$$O_{1A} \quad O_{1B} \quad X \quad O_{2A} \quad O_{2B}$$

where O_{1A} is the pretest observation of construct A, O_{1B} is the pretest observation of construct B, and O_{2A} and O_{2B} are the parallel posttest measures. That the O_{1A} and O_{1B} observations have the same numerical subscripts indicates that they could be collected at the same point in time. The same holds for the O_{2A} and O_{2B} observations. Constructs A and B are chosen so that only construct A is expected to be influenced by the treatment, while both constructs A and B are expected to be influenced by the same threats to internal validity. So, a pretest–posttest difference on construct A accompanied by no pretest–posttest difference on construct B would be evidence in favor of a treatment effect free of the effects of the shared threat to internal validity. For example, Reynolds and West (1987) assessed the effects of an advertising campaign to increase sales of lottery tickets at convenience stores. The sales of lottery tickets (construct A) were compared to the sales of gasoline, cigarettes, and groceries (constructs B) at the same

convenience stores. The advertising campaign was expected to influence sales of lottery tickets but not the other items. But most threats to internal validity, such as those due to history, maturation, instrumentation, testing, and selection differences, were expected to affect measures of both constructs A and B equivalently. That the sale of lottery tickets exhibited an increase from pretest to posttest while the sales of the other items remained constant increased the researchers' confidence that the increase in the sale of lottery tickets was due to the treatment rather than to threats to internal validity.

6.5 STRENGTHS AND WEAKNESSES

The pretest–posttest design is often easy to implement, and the results are easy for even laypersons to understand. The problem with the design results from the plausibility of threats to internal validity. In most research settings, the pretest–posttest design is susceptible to one or more of the threats to internal validity described in this chapter. In estimating the effects of a legislated change of speed limits on freeways, for example, more than one threat to internal validity might be present. Historical events such as changes in the price of gasoline, improvements in automobile safety, or increased vigilance by police for driving under the influence of alcohol or drugs could bias results. So, too, could differences in weather from year to year or season to season. Such threats to internal validity can be just as plausible as a treatment effect as an explanation for any observed difference from pretest to posttest. The effects of threats to internal validity could be as large if not larger, than the effect of the treatment. Because of threats to internal validity, estimates of treatment effects could even be the opposite of the true effect—negative instead of positive, or vice versa.

But the pretest–posttest design will not be biased by threats to internal validity in all research settings, as Eckert (2000) has documented. Eckert (2000) described a simple pretest–posttest design used to assess an educational program where the pretest and posttest were measures of material covered in the program. The educational program lasted less than 2 weeks, during which time there were no historical events that could plausibly have accounted for the observed improvement in test scores. The time interval between pretest and posttest was also too short for significant maturation to occur. An effect due to testing was unlikely because there was nothing about the pretest that could plausibly have influenced the posttest. Instrumentation was not a threat because the pretest and posttest were equivalent measures (with question items assigned randomly to each test). There were no selection differences between pretest and posttest because the composition of the sample was the same at both times. No cyclical changes of significance were present. Regression toward the mean could not account for the results because the participants did not have pretest scores that were selected to be either high or low. Also, the differences from pretest to posttest were statistically significant across the 33 people in the study, so chance was not likely an alternative explanation. The point is that alternative explanations for results of a study threaten the internal validity

of the results only to the extent that alternative explanations are plausible. In some cases, no threats to the internal validity of the pretest–posttest design are plausible.

The results of Eckert's study (which showed a positive effect of the intervention) were credible largely because the time interval between pretest and posttest was short; because there was no plausible way that participants could learn the material being taught during this short interval except by attending the educational intervention; and because the pretest and posttest measures assessed only the materials being taught. When circumstances such as these arise, the pretest–posttest design can produce credible results. However, the researcher must be careful to consider each common threat to internal validity to make sure one or more do not provide plausible alternative explanations for the results.

In some cases, threats to internal validity can be ruled out by collecting auxiliary information. For example, the plausibility that history affects the outcomes can often be assessed by consulting those who are familiar with the research setting or by checking newspapers for reports of potential external influences. In addition, sometimes the basic pretest–posttest design can be strengthened by adding design features such as double pretests or nonequivalent dependent variables.

6.6 CONCLUSIONS

The pretest–posttest design is often easy to implement but is generally a weak quasi-experimental design because threats to internal validity are often plausible as explanations for the results. In those instances in which threats to internal validity are implausible, the design can be used to good effect. But those circumstances are relatively rare when the design is used by itself. Alternative designs are generally preferable.

6.7 SUGGESTED READING

Campbell, D. T., & Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review, 3,* 33–53.
—Discusses threats to internal validity to the pretest–posttest design in the context of a detailed example.

Eckert, W. A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation, 21,* 185–193.
—Explains how a pretest–posttest design can, under the right circumstances, avoid threats to internal validity.

7

Nonequivalent Group Designs

If randomization is absent, it is virtually impossible in many practical circumstances to be convinced that the estimates of the effects of treatments are in fact unbiased.

—Cochran and Rubin (1973, p. 417)

Many [nonequivalent group designs] do not succeed in providing tangible, enduring and convincing evidence about the effects caused by treatments, and those that do succeed often exhibit great care in their design.

—Rosenbaum (1999, p. 259)

Although nonequivalent comparison group studies are quite susceptible to bias, the ability to extract useful information is especially important because many interventions are not amenable to study through an experimental design or a well-controlled quasi-experiment.

—May (2012, p. 507)

Overview

In the nonequivalent group design, groups of participants are nonrandomly assigned to receive either a treatment or a comparison condition. A primary threat to internal validity is the selection differences that are present between the treatment groups. A wide variety of statistical methods have been proposed for analyzing data from nonequivalent group designs to take account of the potentially biasing effects of selection differences. All the statistical methods require strong and often untestable assumptions.

7.1 INTRODUCTION

The nonequivalent group design is one of the most commonly used quasi-experimental designs in field research in the behavioral and social sciences. The design also receives a great deal of attention in the statistics literature where it is usually called an observational study (Cochran, 1965, 1969; Rosenbaum, 2002, 2010). In the Campbellian

tradition, the design is called either the nonequivalent control group design or the nonequivalent comparison group design. I drop "control" and "comparison" from the label to include both cases.

The nonequivalent group design is akin to a between-groups randomized experiment in that both designs draw comparisons between participants who receive different treatment conditions. The difference lies in how participants are assigned to the treatment conditions, whether randomly or nonrandomly. Unlike in between-groups randomized experiments, in nonequivalent group designs, participants are assigned to (or choose) treatment conditions nonrandomly. Nonrandom assignment can take many forms. For example, participants might be assigned nonrandomly by administrators who allocate students to classrooms, patients to hospital wards, or employees to work projects, based on subjective or other nonrandom criteria. Nonrandom assignment also arises when participants select treatments for themselves. For example, participants might choose or not choose to enter smoking cessation programs, job training programs, or psychotherapy treatments based on nonrandom criteria such as level of self-interest, motivation, or desperation.

Following the administration of the different treatments in a nonequivalent group design, the participants are assessed on an outcome (i.e., posttreatment) measure. The difference in outcomes between the nonequivalent groups is then used to estimate the relative effectiveness of the different treatments.

Because of the nonrandom assignment of participants to treatments, participants in the treatment condition likely differ systematically from participants in the comparison condition even before the treatments are implemented. For example, as suggested above, those who choose to participate in a presumed ameliorative treatment might be more self-interested, motivated, or desperate on average than those who choose not to participate in the treatment. As noted in Chapter 4 on randomized experiments, differences between participants in different treatment conditions that exist before the treatments are implemented are called selection differences. Selection differences are a threat to internal validity in nonequivalent group designs. Selection differences might influence the outcomes observed on a posttreatment measure, which could bias the estimate of the treatment effect. That is, participants in the treatment condition might perform differently on the posttest compared to participants in a comparison condition because of initial differences between participants in the treatment conditions, even in the absence of a treatment effect.

If the effect of the treatment is to be accurately estimated, its effects must be disentangled from the effects of selection differences. Otherwise selection differences can bias the estimates of the treatment effect, either masking or masquerading as a treatment effect. For example, the Coalition for Evidence-Based Policy (2003, p. 4) reported that dozens of nonequivalent group studies had "found hormone replacement therapy for postmenopausal women to be effective in reducing the women's risk of coronary heart disease, by about 35–50 percent." But two well-conducted, large-sample randomized experiments found hormone therapy to have the opposite effect: "it *increased* the

risk of heart disease, as well as stroke and breast cancer" (emphasis in original). The reason for the differences in outcomes was that the effects of selection differences were not properly taken into account in the nonequivalent group designs.

7.2 TWO BASIC NONEQUIVALENT GROUP DESIGNS

The simplest nonequivalent group design is the posttest-only nonequivalent group design which is diagrammed thusly:

NR: X O

NR: O

In this design, one group of participants is given the treatment (denoted by the X) and is then measured on a posttest (denoted by the O). A comparison group of participants receives no treatment or an alternative treatment and is then measured, at the same time, on the posttest. The "NR:" at the start of each row and the dashed line between the two rows indicate that the two groups of participants are assigned to treatment conditions nonrandomly. (As noted in Chapter 6, if the two groups were cohorts, the dashed line would be replaced with a dotted line. A dotted line is used with cohorts to indicate that cohorts are likely to be more similar to each other than are other types of nonequivalent groups.) Notice that no observations are collected before the treatment is implemented.

For an example, Mosca and Howard (1997) used this design to assess the effects of grounded learning on outcomes in a business course. The treatment consisted of business executives being brought into the classroom (1) to describe real-life problems to be solved by the students and (2) to provide access to those trying to solve the problems in real-life business settings. The treatment group significantly outperformed the comparison group on multiple measures, including quality of case reports and course grades.

Although this very simple design has been used to produce credible results in some circumstances, the problem with the design is that it provides no way to assess the presence of selection difference, much less take account of the potentially biasing effects of these differences. The design might prove useful in those cases where the researcher is confident the groups differ little at the start and would perform similarly on the posttest were there no treatment effects. Under those conditions, the researcher might feel confident in attributing differences between the groups in the outcome measure to the effect of the treatment. However, critics might not be as willing as the researcher to make such assumptions. So, results from the posttest-only nonequivalent group design typically enjoy little credibility. Researchers who use this design run the risk of producing results that are unconvincing to others, if not themselves.

For these reasons, a slightly more elaborate nonequivalent group design is more common than the simpler (posttest-only) design just presented. The supplemented design adds a pretest and is therefore called the pretest–posttest nonequivalent group design. Such a design is diagrammed thusly:

$$\begin{array}{lccc} \text{NR:} & O_1 & X & O_2 \\ \hline \text{NR:} & O_1 & & O_2 \end{array}$$

In this design, each of the two groups of participants is assessed on a pretest measure. The treatment group then receives the treatment while the comparison group receives either no treatment or an alternative treatment. Then both groups are assessed on the outcome measure—the posttest. Note that although I talk of pretest and posttest measures as if there were only one of each, multiple pretest or posttest measures are possible. That is, the pretest and the posttest could be a battery of measures or observations. My discussion applies regardless of whether there are one or many measures. I talk of *the* pretest and posttest measures simply for convenience, but, at times, I will explicitly talk of multiple measures to emphasize the possibility of a battery of measures.

The pretest measure (or battery of measures) is used both to assess the nature of initial selection differences between the groups and to take account of the effects of these selection differences. (Assessing the nature of initial selection differences should be one of the first steps in the analysis of data from the pretest–posttest nonequivalent group design.) The effect of the treatment is estimated as the difference between the groups on the posttest measure after the pretest measures are used to take account of the effects of selection differences. To be able to take account of the effects of selection differences, the pretests must be collected before the treatments are implemented, or they must be stable traits (such as demographic characteristics like ethnicity, sex, and age) that cannot be affected by the treatments (Rosenbaum, 1984b; Smith, 1957). The pretest need not be operationally identical to the posttest, but the credibility of the analysis can often be greatly increased if the two measures are operationally identical. Because of the advantages of having a pretest measure, from here on I will be concerned only with the pretest–posttest nonequivalent group design rather than with the posttest-only nonequivalent group design. The designation of nonequivalent group design without a modifier will denote the pretest–posttest nonequivalent group design.

The nonequivalent group design is "one of the most commonly implemented research designs in the social sciences" (May, 2012, p. 489). For example, Card and Krueger (1994) used the design to assess the effects of raising minimum wages on employment opportunities; Pischke (2007) to assess the effects of the length of the school year on academic performance and subsequent earnings; Hong and Raudenbush (2005) to assess the effects of retention in grade level on learning; Rubin (2000) to study of the effects of smoking on health care costs, Langer and Rodin (1976) to assess the

effects of increased opportunities for decision making on health among nursing home residents; Lehman, Lampert, and Nisbett (1988) to assess the effects of coursework in statistics on quantitative reasoning ability; Hackman, Pearce, and Wolfe (1978) to assess the effects of making jobs more complex and challenging on the attitudes and behaviors of workers; Molnar et al. (1999) to assess the effects of reduction in class size on achievement; and West et al. (2014) to assess the effect of telephone counseling on the well-being of sufferers from chronic diseases.

In all these examples, selection differences posed a threat to internal validity. As a result, all these studies addressed the threat of selection differences. There are multiple ways to address selection differences (Schafer & Kang, 2008; Winship & Morgan, 1999), the most common of which are described next. For educational researchers, I might note that nonequivalent group designs with pretest differences greater than .25 of the within-group standard deviations do not meet the design standards of the What Works Clearinghouse, regardless of the statistical procedures used to take account of the potentially biasing effects of selection differences (U.S. Department of Education, 2017).

7.3 CHANGE-SCORE ANALYSIS

A **change-score analysis** requires a pretest measure that is operationally identical to the posttest measure (Allison, 1990). For example, if the posttest is a measure of mathematical ability, the pretest must be the same or a parallel measure of mathematical ability. Or if income is the posttest measure, income must also be the pretest measure. In addition, credibility is increased with high correlations between the pretest and posttest. For example, the What Works Clearinghouse requires that the pretest–posttest correlation must be .6 or higher if a change-score analysis is to be an acceptable method for taking account of initial selection differences (U.S. Department of Education, 2017).

The change-score analysis is conducted by subtracting the pretest scores from the posttest scores and entering this difference (or change) score in an ANOVA model:

$$(Y_i - X_i) = \alpha + (\beta_T T_i) + \varepsilon_i \tag{7.1}$$

where Y_i is the ith participant's posttest score, X_i is the ith participant's pretest score, and T_i is an indicator variable representing the treatment assignment for the ith participant, where T_i equals 0 if the participant is assigned to the comparison condition and T_i equals 1 if the participant is assigned to the treatment condition. The dependent variable is the change score $(Y_i - X_i)$, which is the difference between the posttest and pretest scores for each participant. The estimate of α is equal to the mean change score for the participants in the comparison condition. The estimate of β_T is the regression coefficient for the T_i variable and is equal to the mean difference in change scores

between the treatment and comparison groups. The estimate of β_T is the estimate of the treatment effect. The ε_i term is the residual and represents the unexplained variance in the change scores—that is, the variance in the change scores that is not explained by the rest of the model.

In the change-score analysis, the null hypothesis is that, in the absence of a treatment effect, the average change from pretest to posttest would be the same for the treatment and comparison groups. This is called the common trends or **parallel trends assumption.** An equivalent way to express the same null hypothesis is that, in the absence of a treatment effect, the average pretest difference between the treatment groups would be the same as the average posttest difference between the treatment groups. This is the way the change-score analysis takes account of selection differences—by assuming that the effects of selection differences are the same on the pretest as on the posttest. The same results as the change-score analysis would be obtained with a repeated-measures analysis of variance. The change score analysis is also called a **difference-in-differences (DID) analysis** (see also Section 9.8 for more elaborate difference-in-differences analyses).

The parallel trend assumption is very restrictive. The parallel trend assumption (and hence the change-score analysis) might be appropriate in some cases but certainly not in all. For the parallel trend assumption to be plausible, the pretest and posttest observations will generally need to be at least interval-level measures (where interval-level measurement means that a 10-point difference, for example, at one location on the pretest and posttest scale is the same as a 10-point difference at all other locations—which is the case with measures such as the Fahrenheit measurement of temperature). If the pretest and posttest are not both interval-level measures and if the groups start out at different levels on the pretest, in the absence of a treatment effect, there may be little reason to believe that average growth would be equivalent in the two groups. Even if the pretest and posttest are at least interval-level measures in theory, if the groups start out at different locations on the pretest score, floor or ceiling effects can make average growth unequal in practice, even in the absence of a treatment effect. And even if the pretest and posttest are at least interval measurements and there are no floor or ceiling effects, the two treatment groups may have different natural rates of growth in the absence of a treatment effect, which would invalidate the null hypothesis. If the treatment and comparison groups start at different levels on the pretest, there must be a reason. Whatever the reason, that reason may cause the groups to grow further apart over time, even in the absence of a treatment effect. For example, the "rich getting richer" is a metaphor for groups growing further apart over time even in the absence of an external intervention such as a treatment. Consider pretests and posttests that assess mathematical ability. One treatment group may perform better on the pretest than the other group because the higher-performing group is better at mathematics than the other group. This initial advantage may well grow over time to produce an increasing gap between the groups in mathematical performance, even in the absence of any treatment effect. Such differential growth rates are said to result in a differential maturation

or selection-by-maturation threat to internal validity. Such a threat can be addressed by other methods of analysis described below, but a selection-by-maturation threat seriously undermines a change-score analysis.

Alternatively, differential regression toward the mean can cause the treatment groups to move closer together over time. For example, participants might be placed into the treatment condition because they scored below average on the pretest and participants might be placed into the comparison condition because they scored above average on the pretest. May (2012) argues that this could be the case in studies of antidepressant medications where those prescribed the medications are those who are most severely depressed initially. In the presence of such initial selection differences, the groups would be expected to move closer together over time simply because of differential regression toward the mean (Campbell & Erlebacher, 1970; Campbell & Kenny, 1999). That is, the below-average pretest scores for participants in the treatment condition should increase by the time of the posttest measurement to be closer to the average, while the above-average pretest scores for the participants in the comparison condition should decrease by the time of the posttest measurement to be closer to the average. As a result, differential regression toward the mean can bias the estimate of the treatment effect in a change-score analysis.

The point here is that bias can arise in numerous ways in a change-score analysis. The change-score analysis is likely to be most appropriate under either of two sets of conditions (Rosenbaum, 2017; Shadish et al., 2002). The first set of conditions is that the treatment groups start out at the same level on the pretest measure, and there is no reason to believe the groups would grow at different rates over time or regress toward different means, in the absence of a treatment effect. Assuming the treatment will raise (lower) scores, the second set of conditions is that the treatment group starts out below (above) the comparison group on the pretest and ends up above (below) the comparison group on the posttest. This would result in a crossover interaction in the data, which is usually difficult to explain by causes such as the rich getting richer or regression toward the mean (for examples, see Rosenbaum, 2005a; Wortman, Reichardt, & St. Pierre, 1978). Such crossover interactions are rare, however, because they require a sufficiently strong treatment effect to overcome the effects of initial selection difference, and such strong treatment effects are relatively rare. The change-score analysis has its place in the arsenal of statistical methods used to analyze data from the nonequivalent group design. But its appropriate use is usually limited by the research circumstances.

A statistical analysis that is equivalent to the change-score analysis is also possible. Instead of calculating change scores, stack the pretest data on top of the posttest data. For example, if there were 10 participants in the study, there would be 20 rows of scores in the data matrix, with the first 10 being the pretest data and the last 10 the posttest data. The following model could then be fit:

$$\begin{aligned} Y_{ij} &= (\beta_1 \, \text{TIME}_{ij}) + (\zeta_1 \, P_{1ij}) + (\zeta_2 \, P_{2ij}) + \ldots + (\zeta_N \, P_{Nij}) \\ &\quad + [\beta_2 \, (\text{TIME}_{ij} \times \text{CONDITION}_{ij})] + \varepsilon_{ij} \end{aligned} \tag{7.2}$$

where

i represents the ith participant (out of a total of N);
j represents time of measurement: equals 0 for the pretest scores and 1 for the posttest scores;
Y_{ij} is the ith participant's pretest score if $j = 0$ and the posttest score if $j = 1$ (so there is a total of $2N$ entries for Y_{ij});
TIME_{ij} is an indicator variable scored 0 for pretest data and 1 for posttest data;
P_{1ij} is an indicator variable coded 1 for the first participant and 0 otherwise;
P_{2ij} is an indicator variable coded 1 for the second participant and 0 otherwise;
. . .
P_{Nij} is an indicator variable coded 1 for the Nth participant and 0 otherwise (so there are as many P variables as there are participants);
CONDITION_{ij} is an indicator variable scored 0 for participants in the comparison condition and 1 for participants in the treatment condition;
$\text{TIME}_{ij} \times \text{CONDITION}_{ij}$ is the product of TIME_{ij} and CONDITION_{ij}; and
ε_{ij} is the residual.

Notice that no α (i.e., intercept) term is included in the model (although an intercept can be included using an alternative parameterization by omitting one of the P variables to avoid perfect multicollinearity). The values of ζ_1 to ζ_N are effects for each participant. The treatment effect estimate is the estimate of β_2. Equation 7.2 is often called a DID model.

I might note that the most basic DID analysis uses a simpler specification with a single CONDITION_{ij} variable in place of the multiple P variables in Equation 7.2. The simpler model is written

$$Y_{ij} = \alpha + (\beta_1 \text{ TIME}_{ij}) + (\beta_2 \text{ CONDITION}_{ij}) + [\beta_3 (\text{TIME}_{ij} \times \text{CONDITION}_{ij})] + \varepsilon_{ij} \quad (7.3)$$

where the variables and subscripts are defined as in previous equations. The treatment effect estimate is the estimate of β_3. This value will be the same as in the models given in Equations 7.1 and 7.2. But the standard error for the treatment effect from Equation 7.3 will not be correct if the model is fit using ordinary least squares (OLS). To get a correct standard error you would have to take account of the relationships within participants between their pretest and posttest scores (Bertrand, Duflo, & Mullainathan, 2004). Wing, Simon, and Bello-Gomez (2018), along with Angrist and Pischke (2015), suggest making the adjustment by using clustered standard errors.

I present Equation 7.3 not to endorse it but simply to note that it is widely reported in the DID literature as an appropriate model for the given data. I believe, however, that the other two equations (7.1 and 7.2) are superior because they do not require an additional adjustment to the standard error.

7.4 ANALYSIS OF COVARIANCE

The analysis of covariance (ANCOVA) was introduced in Section 4.6 in the chapter on randomized experiments. The statistical model remains the same regardless of whether the design is a randomized experiment or nonequivalent group design. As before, the basic ANCOVA model is

$$Y_i = \alpha + (\beta_T T_i) + (\beta_X X_i) + \varepsilon_i \tag{7.4}$$

where Y_i represents scores on the posttreatment observation, T_i is an indicator variable representing the treatment assignment for the ith participant (as described before), X_i represents the scores on the pretreatment observations, and ε_i is the residual (see Equation 4.2 in Section 4.6.1). The estimate of β_X is the estimated slope of Y regressed on X within each treatment condition where the two regression lines are constrained to have equal slopes. The estimate of β_T is the estimate of the treatment effect.

Although the ANCOVA equations are the same in the nonequivalent group design and the randomized experiment, ANCOVA functions differently in the two designs, as can be seen from graphs of data. In the randomized experiment, initial selection differences between the treatment groups are random. As a result, the scatters of data points from the two treatment groups lie above one another in a scatterplot of the posttest on the pretest scores, with any horizontal displacement of the scatters due only to random selection differences (see Figures 4.3 and 4.4). In the nonequivalent group design, the two scatters of points can be displaced horizontally as in Figure 7.1. That is, the pretest scores in the treatment group can differ systematically from the pretest scores in the comparison group—so that $\overline{X}_T$ can differ systematically from $\overline{X}_C$—as revealed in Figure 7.1 (where $\overline{X}_T$ and $\overline{X}_C$ are the pretest means in the treatment and comparison groups, respectively). In Figure 7.1 (as in Figures 4.3 and 4.4), the estimate of the treatment effect from the ANCOVA model (i.e., the estimate of β_T) is the vertical displacement of the regression lines, which is also the difference between the intercepts (labeled α_T and α_C for the treatment and comparison groups, respectively, in Figure 7.1) of the two regression lines. The difference between the intercepts can deviate substantially from the difference between the posttest means in the treatment and comparison groups, which are labeled $\overline{Y}_T$ and $\overline{Y}_C$, respectively, in the figure, as explained below.

In the randomized experiment, the ANOVA model (see Equation 4.1 in Section 4.5) and the basic ANCOVA model (see Equation 4.2 in Section 4.6.1) are estimating the same treatment effect. That is, the expected value of the estimate of the treatment effect is the same in the two analyses. Because the ANCOVA adds a covariate to the analysis compared to ANOVA, the ANCOVA makes an adjustment for random selection differences, while the ANOVA does not. Even so, this does not change the expected value of the estimate of the treatment effect in the randomized experiment. The ANOVA estimate is the difference between the posttest means in the two groups ($\overline{Y}_T - \overline{Y}_C$; see Section 4.5). The ANCOVA estimate is the vertical displacement between the regression

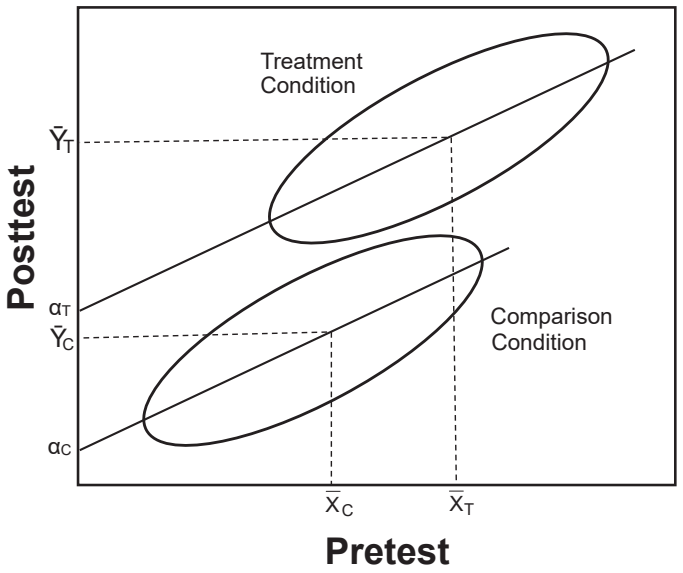


FIGURE 7.1. Idealized scatterplots from a nonequivalent group design. In the nonequivalent group design, the two treatment conditions generally do not have the same pretest means, and the difference between the posttest means ($\bar{Y}_T$ and $\bar{Y}_C$) is not generally the same as the difference between the intercepts (α_T and α_C).

lines in the two groups, which is equal to the mean posttest difference adjusted for the mean difference in the pretests $[(\bar{Y}_T - \bar{Y}_C) - B_X(\bar{X}_T - \bar{X}_C)]$, where B_X is the estimate of the regression slopes β_X (see Section 4.6.1). But because the scatters of the two groups are displaced horizontally in the randomized experiment only by chance differences (i.e., because $(\bar{X}_T - \bar{X}_C)$ equals zero in expected value), the treatment effect estimates from the ANOVA and ANCOVA are equal in expected value.

In contrast, in the nonequivalent group design, the expected values of the treatment effect estimates in the ANOVA and ANCOVA are generally not the same. The difference in expected values between the ANOVA and ANCOVA estimates in the nonequivalent group design arises because the ANCOVA makes a correction for the presence of systematic nonrandom selection differences in the pretests, while the ANOVA does not. That is, the ANOVA and ANCOVA estimates are not the same in expected value in the nonequivalent group design because nonrandom selection differences systematically displace the scatters of data points in the two treatment groups horizontally (i.e., because $(\bar{X}_T - \bar{X}_C)$ is not equal to zero in expected value). Figure 7.2 reveals how the treatment effect estimates in the ANOVA and ANCOVA can differ in the nonequivalent group design. In all the panels in Figure 7.2, the ANOVA estimate is the mean difference between the treatment groups on the posttest scores, which is the difference between $\bar{Y}_T$ (the mean posttest score of the treatment group) and $\bar{Y}_C$ (the mean posttest score for the comparison group). In the ANCOVA, the estimate of the treatment effect is the vertical displacement between the regression line in the treatment group and the regression

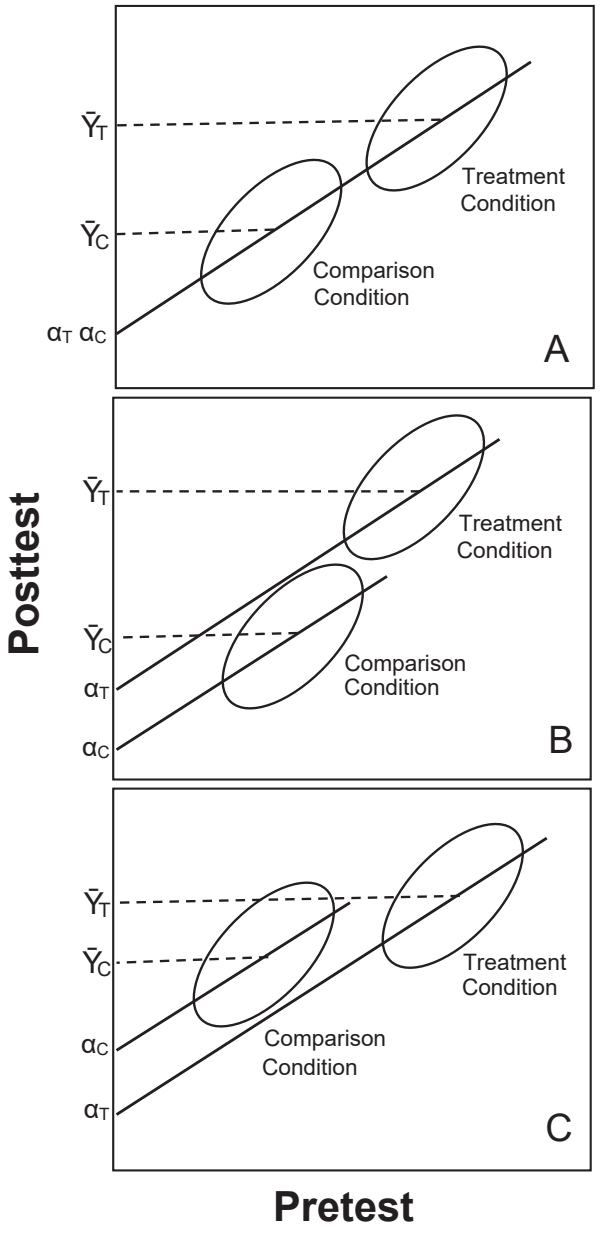


FIGURE 7.2. How the treatment effect estimates from ANOVA and ANCOVA differ in the nonequivalent group design. In all three panels, the posttest mean in the treatment condition ($\bar{Y}_T$) is larger than the posttest mean in the comparison condition ($\bar{Y}_C$), so ANOVA would estimate the treatment effect to be positive. Panel A: ANCOVA would estimate the treatment effect to be zero because the intercepts from the two regression lines (α_T and α_C) are equal. Panel B: ANCOVA would estimate the treatment effect to be positive (but smaller than the estimate from ANOVA) because the intercept in the treatment condition (α_T) is larger than the intercept in the comparison condition (α_C), but the difference between intercepts is smaller than the difference between posttest means. Panel C: ANCOVA would estimate the treatment effect to be negative because the intercept in the treatment condition (α_T) is smaller than the intercept in the comparison condition (α_C).

line in the comparison group, which is equal to the difference in the two regression line intercepts (see Section 4.6.1). In Panel A in Figure 7.2, because the mean posttest score from the treatment group is larger than the mean posttest score for the comparison group, the estimate of the treatment effect in the ANOVA would be positive. This estimate is different from the estimate of the treatment effect in the ANCOVA. Because the two regression lines fall on top of each other, they have the same *Y* intercepts (denoted by α_T and α_C in the figure), and hence the ANCOVA would estimate the treatment effect to be zero.

Things are different in Panel B in Figure 7.2. Here the ANOVA would still find that the treatment had a positive effect because $\overline{Y}_T$ is greater than $\overline{Y}_C$. The ANCOVA would also find that the treatment had a positive effect, but the estimate from the ANCOVA would be smaller than the estimate from the ANOVA, as can be seen by comparing the difference between $\overline{Y}_T$ and $\overline{Y}_C$ and the difference between the *Y* intercepts of the two regression lines (α_T and α_C). Things are different once again in Panel C in Figure 7.2. Here the ANOVA would again find that the treatment had a positive effect (as can be seen from the difference between $\overline{Y}_T$ and $\overline{Y}_C$). In contrast, the ANCOVA would find the treatment effect to be negative (as can be seen from the difference between α_T and α_C). As these panels suggest, the ANOVA and ANCOVA estimates can differ in all possible ways. One treatment effect estimate can be large and the other small. Both estimates can be in the same direction, or one can be positive while the other is negative. The difference between the ANOVA and ANCOVA estimate arises, to repeat, because the ANCOVA takes account of selection differences on the pretest scores while the ANOVA does not. This can be seen algebraically, as well as pictorially, from the treatment effect estimates for the two analyses. As noted above, the treatment effect estimate from ANOVA is $(\overline{Y}_T - \overline{Y}_C)$, while the estimate of the treatment effect from ANCOVA is $[(\overline{Y}_T - \overline{Y}_C) - B_X (\overline{X}_T - \overline{X}_C)]$ (again see Sections 4.5 and 4.6.1). Because the average selection differences on the pretest, $(\overline{X}_T - \overline{X}_C)$ can be nonzero in expected value in the nonequivalent group design, the two treatment effect estimates can differ in expected value in the nonequivalent group design.

Alternatively, the difference between ANOVA and ANCOVA can be conceptualized as a difference in matching. ANCOVA makes an adjustment for selection differences by mathematically matching participants from the two groups on the pretest. That is, ANCOVA estimates the treatment effect by comparing posttest outcomes from participants in the two treatment groups who have first been matched on their pretest scores (see Section 4.6.1). So the vertical difference between the regression lines (and hence the difference in intercepts) is equal to the mean difference in posttest scores for participants matched on their pretest scores. ANOVA compares the posttest outcomes from participants in the two treatment groups without any matching on the pretest scores.

Perhaps it is also worth commenting on the difference between the change-score analysis and the ANCOVA in the nonequivalent group design (also see Reichardt, 1979). As just noted, the estimate of the treatment effect in the ANCOVA is $[(\overline{Y}_T - \overline{Y}_C) - B_X (\overline{X}_T - \overline{X}_C)]$. The treatment effect estimate in the change-score analysis is the same except that

the value of B_X is set to 1 and so the treatment effect estimate becomes $[(\overline{Y}_T - \overline{Y}_C) - (\overline{X}_T - \overline{X}_C)]$. As just explained, including the pretest as a covariate in the ANCOVA estimate makes the ANCOVA equivalent to a matching strategy where participants are compared on their posttest scores after being matched on their pretest scores. The change-score analysis matches in a different fashion. As explained earlier, the change-score analysis specifies that the treatment conditions are matched on their change over time. That is, the treatment conditions are expected, under the null hypothesis, to have the same average change over time—from pretest to posttest. The ANCOVA does not impose this assumption, but by matching on the pretest scores, it allows the average expected change over time to vary across the treatment conditions (because B_X need not be equal to 1).

Just as in the randomized experiment, interaction terms can be added to the ANCOVA model in the nonequivalent group design, if the regression slopes in the two treatment conditions differ (see Equation 4.4 in Section 4.6.2). Assuming there is no **hidden bias** (which is a major assumption; see Section 7.4.1), if the pretest scores (X_i) are centered at the overall mean of X_i in the presence of an interaction (as in Equation 4.4 in Section 4.6.2), the ANCOVA model would estimate the average treatment effect (ATE) across all participants (which is the effect you would get by comparing the outcome obtained if all participants received the treatment condition to the outcome obtained if all participants received the comparison condition) (Aiken & West, 1991; Schafer & Kang, 2008). In contrast, if the pretest scores are centered at the mean of the pretest scores in the treatment condition (i.e., centered at $\overline{X}_T$) the ANCOVA model would estimate the average treatment effect on the treated (called the ATT which is the effect you would get were you to compare the participants in the treatment group if they received the treatment to the participants in the treatment group if they received the comparison condition) (Austin, 2011; Cochran, 1957; Schafer & Kang, 2008). Without interactions in the linear ANCOVA model, the ATE and ATT estimates are the same (and are both equal to the vertical displacement between the regression lines). When the two estimates differ, circumstances dictate whether the ATE or the ATT estimate is most useful. The ATT estimate is most relevant when it makes most sense to generalize only to those participants who have been willing to accept or seek out the treatment assuming the treatment is likely to be given in the future only to those who are willing to accept or seek it out. On the other hand, the ATE estimate is likely to be most relevant if future implementations of the treatment are likely to include the full range of participants in the study.

Just as in the randomized experiment, when using ANCOVA, it is important in the nonequivalent group design to correctly model the shape of the regression surface between posttest and pretest in each treatment group. As in the randomized experiment, polynomial terms can be added to the ANCOVA model in the nonequivalent group design to take account of curvilinear relationships between the pretest and posttest (see Section 4.6.3). In the presence of nonlinear regression surfaces and interactions, however, no parameter in the model equals the average treatment effect (either

the ATE or the ATT). To calculate either the ATE or the ATT, respectively, the researcher would have to estimate the treatment effect at each value of X_i and average these values across either all participants or just those participants in the treatment condition (Schafer & Kang, 2008). Alternatively, nonlinear regression surfaces could be fit using procedures such as loess or spline regression. The same procedures as just mentioned would have to be used to estimate either the ATE or the ATT.

Additional covariates can be added to the ANCOVA model, so selection differences on all the covariates are taken into account. In this case, the ANCOVA would estimate the treatment effect as the posttest difference between participants who were matched on all the included covariates. In other words, the statistical equating that ANCOVA performs is much the same as would be accomplished if participants from the two treatment groups were physically matched on their scores on all the included covariates and then compared on their posttest scores (assuming the regression surfaces have been fit correctly in the ANCOVA model).

7.4.1 Hidden Bias

Just because the ANCOVA matches participants from the two treatment groups on the covariates entered in the model does not mean the matching is sufficient to remove all bias due to selection differences. Hidden bias is the bias, due to selection differences, that remains after the entered covariates have been taken into account. That is, hidden bias arises from selection differences on covariates not included in the statistical analyses because all the covariates on which there are selection differences might not be observed.

To completely remove bias due to selection differences (called **selection bias**), the covariates included in the model must well model either the posttest or the selection differences (Austin, 2011; Barnow, Cain, & Goldberger, 1980; Cronbach, Rogosa, Floden, & Price, 1976, 1977; Reichardt, 1979) or both together (which is the approach taken, for example, in **doubly robust methods**—Funk et al., 2011; Schafer & Kang, 2008). To model the posttest (and thereby remove all bias due to the effects of selection differences), the covariates must mirror the causal forces operating on the posttest—except for the effect of the treatment and any causal forces unrelated to the selection process. Alternatively, to model selection differences (and thereby remove all bias due to the effects of selection differences), the covariates must mirror the causal forces by which participants were selected into treatment groups—except for any causal factors that are unrelated to the posttest. If either of these conditions holds, the treatment assignment is ignorable, conditional on the covariate measures (Rosenbaum, 1984a; Rosenbaum & Rubin, 1984). **Ignorability** is also called **unconfoundedness,** conditional independence, absence of hidden bias, selection on observables, or absence of omitted variable bias. This is just another way of saying that equating the treatment groups on the covariates is sufficient to remove bias due to selection differences. (**Strong ignorability** arises if, in addition to ignorability, each participant has a nonzero probability of being in either the treatment or comparison condition.)

Unfortunately, however, neither of the two conditions for ignorability might hold true. Consider each in turn. First, the covariates might not well model all the causal forces operating on the posttest (that are related to selection). For example, the pretest is likely to be the single best predictor of the posttest and, among all covariates, is the measure most likely to have the same causal structure as the posttest. Different causal forces may nonetheless be operating on the pretest than on the posttest because the structure of causal forces changes over time (which is one reason why the pretest–posttest correlation is seldom perfect). In this case, the pretest would not provide an adequate causal model of the posttest. Or even if the causal forces are the same in the pretest and posttest, they might not be operating in the same proportion in the two measures. Another way to say the same thing is that the loadings in the factor structure of the pretest may not be the same as the loadings in the factor structure of the posttest (Pitts et al., 1996; Reichardt, 1979). For example, the weightings of different mathematical skills (arithmetic versus algebra) for scores on a pretest might not be the same as the weighting for scores on the posttest, even if the tests remain the same. Including other variables as covariates might still not provide a proper causal model for the posttest. For example, suppose the treatment groups are growing at different rates in the absence of a treatment effect. Then the covariates must predict the different growth rates if they are to provide an adequate causal model of the posttest (Haviland, Nagin, & Rosenbaum, 2007). But the available covariates might not be sufficiently prescient to accomplish that task.

Second, the covariates (including the pretest) might not adequately model selection differences because, for example, the covariates do not well account for all the factors that contribute to selection and that are related to the outcome. For example, selection into the treatment groups might be largely due to motivation (and motivation might influence the outcome scores), but all the relevant individual differences in motivation might not be well captured by either the pretest or the other covariates.

To recap the discussion, consider the example given by May (2012, p. 491) about the effect of antidepressant medications on depression. Suppose both the treatment and comparison groups receive talk psychotherapy and, in addition, that the treatment group receives antidepressant medications. The substantive question being asked is whether adding antidepressant medications to talk therapy provides any benefits (or has any drawbacks) compared to talk therapy alone. Perhaps talk therapy is more effective when accompanied by antidepressant medications. Or perhaps antidepressant medications interfere with a patient's emotional commitment to talk therapy and so produces worse outcomes than talk therapy by itself. Now assume that the effects of the different treatments are assessed with a nonequivalent group design.

The first question to ask is whether the pretest (and other covariates) properly model the causal structure of the posttest. The answer might be "no" because what causes depression at the time of the pretest (before psychotherapy) may not be the same as what causes depression at the time of the posttest (after psychotherapy). Thus, the pretest and covariates might not provide an adequate model of the later causes.

The second question to ask is whether the pretest (and other covariates) properly model selection differences between the treatment groups. Again, the answer might be "no" because of the many factors that lead to selection into the treatment group. As May (2012) points out, selection into the treatment group might depend on whether participants (1) select physicians who are willing to prescribe antidepressant medications, (2) are seen by their physicians as appropriate for treatment with antidepressant medications, (3) are members of health maintenance organizations that pay for antidepressant medications, (4) are willing to take antidepressant medications, and so on. Unless all these factors have been measured and included as covariates in the ANCOVA model, the analysis might not adequately model the selection differences that distinguish participants in the treatment condition from the participants in the comparison condition.

The point is the following. An ANCOVA model might well be crafted that can adjust for selection differences between the treatment groups in observed covariates. There is no guarantee, however, in the nonequivalent group design that controlling for the observed covariates is adequate to remove all the bias due to selection differences in the estimate of the treatment effect. For example, just controlling for a few demographic measures is not likely to be adequate to model either the outcome scores or the process of selection. ANCOVA will likely be most appropriate for analyzing data from the nonequivalent group design when (1) the covariates (which include a pretest that is operationally identical to the posttest) are highly correlated with the posttest to provide the best possible causal model of the posttest scores, and (2) the covariates well represent the selection mechanisms by which participants are selected into the treatment groups, and (3) efforts have been made to take account of measurement error in the covariates (see Section 7.4.2). Even then, there is no guarantee that all hidden bias has been removed. There is no guarantee that estimates of treatment effect from ANCOVA will be unbiased.

7.4.2 Measurement Error in the Covariates

Although it may not be obvious, measurement error (even random measurement error) in a covariate, including the pretest, can bias the estimate of the treatment effect in an ANCOVA (Cochran, 1968a; Reichardt, 1979). For simplicity, suppose that the only covariate included in the analysis is the pretest. Further suppose that the pretest equals the posttest in the absence of measurement error in either the pretest or the posttest. Then the pretest is a perfect model of the posttest and would completely remove bias due to selection differences in an ANCOVA. Under these conditions, data that illustrate the null result of no treatment effect are shown in Panel A of Figure 7.3. Because the pretest equals the posttest, all the scores fall on top of the regression lines without any scatter. With the assumption of no treatment effect, the Y intercepts of the two regression lines would be the same, so the ANCOVA would provide the correct estimate of the treatment effect, which is zero.

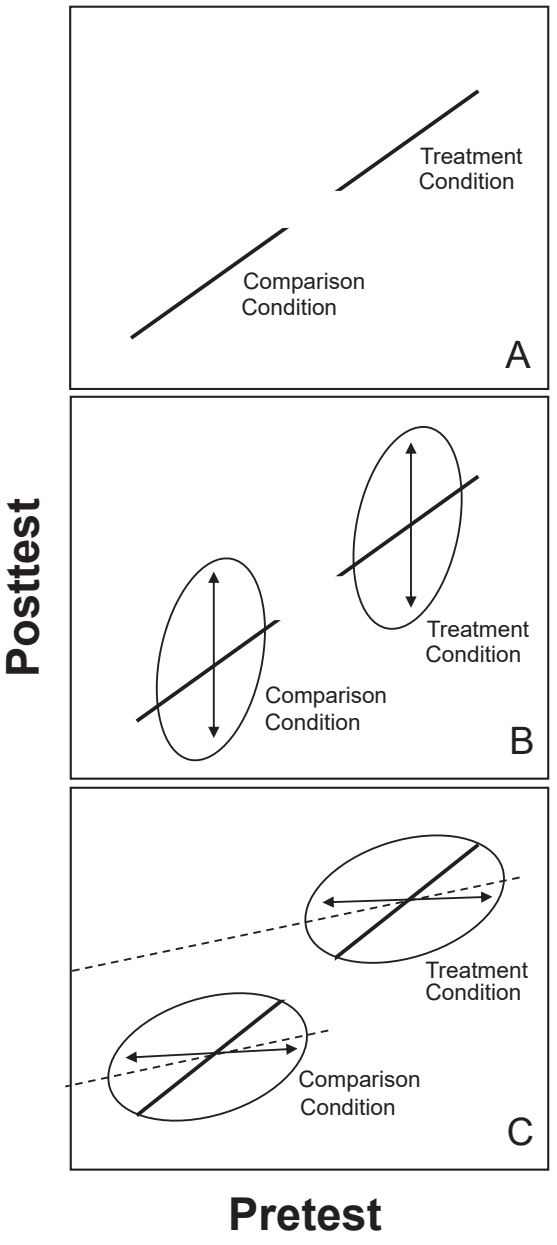


FIGURE 7.3. The effects that measurement error in the pretests and posttests has on the treatment effect estimate in ANCOVA. Each panel reveals a scatterplot for the scores in each of the two treatment conditions. The arrows indicate the presence or absence of measurement error. Panel A: The case of no measurement error and a perfect relationship between pretest and posttest. This is an instance where the treatment is correctly estimated to have no effect. Panel B: Random measurement error is added to the posttest—the treatment is still estimated to have no effect. Panel C: Random measurement error is added to the pretest instead of the posttest. A bias results wherein the treatment is estimated to have an effect.

The same conditions hold in Panel B of Figure 7.3 except that random measurement error has been added to the posttest. The ellipses in the figure reveal how the data would be scattered around the regression lines, with the arrows showing how measurement error causes the scatter. Notice that random measurement error in the posttest causes scatter around the regression lines in the vertical direction but does not alter the regression lines. The regression lines still go through the mean of the posttest scores for a given pretest score (see Section 4.6.1). Because the regression lines are unaltered, the *Y* intercepts for the regression lines in the two treatment groups remain the same. An ANCOVA fit to these data would again provide the correct estimate of the treatment effect, which is zero. The point is that random measurement error in the posttest does not bias the estimate of the treatment effect in an ANCOVA. However, a different conclusion is reached if random measurement error is added to the pretest, as demonstrated next.

In Panel C of Figure 7.3, random measurement error in the posttest has been removed and has been added instead to the pretest. Again, the ellipses in the figure reveal how the data would be scattered around the regression lines, and the arrows show how measurement error causes the scatter. Note how measurement error in the pretest spreads the data out in the horizontal direction. Also note that with random error in the pretest, the regression lines that best fit the scatters are tilted (attenuated), as indicated by the dashed lines in the figure compared to the true regression lines. This tilt results in a vertical displacement between the fitted regression lines, which means that the regression lines that best fit the scatters do not have the same *Y* intercept. As a result, an ANCOVA would estimate the treatment effect to be nonzero, which is incorrect. The lower is the reliability of the pretest, the greater is the tilt in the regression lines, and the less ANCOVA corrects for selection differences on the pretest. (With zero reliability, the regression lines would be horizontal and thereby make no correction at all for selection differences.) The point is that measurement error in the pretest (even when it is random) biases the estimate of the treatment effect in ANCOVA.

The best approach, obviously, is to measure the pretest without error, but this is often difficult, if not impossible. The alternative is to correct for the effects of measurement error in the pretest. Several methods have been proposed for this purpose. The simplest is due to Porter (1967; Porter & Chibucos, 1974). The Porter method requires an estimate of the reliability of the pretest within the treatment groups. (The proper measure of reliability—e.g., internal consistency or test–retest—depends on which sources of error are responsible for making the observed pretest deviate from the underlying construct on which matching should take place.) Assuming the reliability of the pretest is the same in the two treatment groups, we find that the correction is the following. Calculate an adjusted X_i score in the following fashion:

$$X_{\text{adjusted}i} = \overline{X}_j + [p_{XX}\,(X_i - \overline{X}_j)] \tag{7.5}$$

where $X_{\text{adjusted}i}$ is the adjusted pretest score for the *i*th participant, p_{XX} is the estimate of the within-group reliability of the pretest scores, X_i is the pretest score for the *i*th

participant, and $\overline{X}_j$ is the mean of the pretest scores in the ith participant's treatment group (May, 2012; Reichardt, 1979). The treatment effect is then estimated by entering the adjusted pretest scores in place of the original pretest scores in the ANCOVA. This method obviously depends on knowing the reliability of the pretest and is applicable only if the pretest is the only covariate in the ANCOVA model.

An alternative method for addressing measurement error in covariates (including the pretest) is to use **structural equation modeling (SEM),** with the covariates represented by **latent variables** (Bentler & Woodward, 1978; Magidson 1977; Magidson & Sörbom, 1982). Latent variable SEM requires multiple measures for each construct being measured by the covariates (Kline, 2016; Loehlin & Beaujean, 2017). For example, if participants are to be equated on motivation, motivation must be assessed using at least two separate measures (and preferably three or more). In essence, latent variable SEM creates a factor analysis model of each underlying, latent construct (e.g., motivation) using multiple measures for that construct. Also see Aiken, Stein, and Bentler (1994) for an alternative approach to using SEM to account for errors of measurement. However, whether latent variable SEM correctly takes account of the biasing effects of measurement error in the covariates depends on having a correct latent variable model for the measurement error. Even if measurement error is correctly modeled, estimates of treatment effects will still be biased if the pretest measures do not adequately take account of selection differences (i.e., if there is hidden bias). In addition, latent variable models require large sample sizes. An instrumental variables approach (see Section 7.7) can also be used to take account of measurement error in the covariates. But whatever method is used, keep in mind that taking account of measurement error does not take account of hidden bias.

Measurement error in the covariates does not bias the estimate of the average treatment effect when using ANCOVA in the randomized experiment. Measurement error in a covariate still attenuates the regression slopes (see Figure 7.3), but the expected value of the average treatment effect remains unchanged. Recall that the estimate of the average treatment effect in the basic ANCOVA model is $(\overline{Y}_T - \overline{Y}_C) - B_X(\overline{X}_T - \overline{X}_C)$, where B_X is the estimate of β_X in Equation 4.2 (see Section 4.6.1). While the value of B_X varies with the amount of measurement error, the expected value of $(\overline{X}_T - \overline{X}_C)$ is always zero in the randomized experiment. Therefore, the expected value of the average treatment effect estimate remains unchanged even as the value of B_X varies.

7.5 MATCHING AND BLOCKING

An ANCOVA adjusts for selection differences by matching participants mathematically on baseline measures. The matching is performed using estimated regression lines (as shown in Figure 4.3). An alternative is to equate participants in the two treatment groups by physically matching them on baseline measures (Stuart, 2010). That is, participants in the treatment group are physically matched on the covariates with participants in the

comparison group. The treatment effect is then estimated by comparing the outcome scores of the matched participants. Such matching takes account of selection differences in much the same way as ANCOVA does (because ANCOVA is also a matching procedure). It has much the same advantages and limitations, assuming the regression surfaces have been correctly modeled in ANCOVA.

Matching is more complex in the context of the nonequivalent group design than in that of randomized experiment (see Section 4.6.4). When matching is most often used in a randomized experiment, participants are matched on their baseline measures and then randomly assigned to treatment groups. That is, the treatment groups are formed after the participants have been matched. In the nonequivalent group design, groups are already formed, and the researcher has to match participants from the two treatment groups after the fact—and then estimate the treatment effect by comparing the newly matched subsets of participants.

If matching is performed on a single covariate, a common procedure is to choose a **caliper distance** as a criterion for a match. If participants from each treatment group are within the caliper distance on the covariate, they can be matched. If two participants are not within the caliper distance, however, they cannot be matched. If no match can be found for a participant, that participant is dropped from the analysis. Recommendations about the size of the caliper distance vary. Cochran (1972; also see Cochran & Rubin, 1973) showed that a caliper distance of .2 of the within-group standard deviation of a covariate removes about 99% of the bias due to that covariate. Other researchers use a caliper distance of only .1 of the within-group standard deviation on the covariate (Steiner & Cook, 2015). The smaller the caliper distance, the higher the quality of matches and the greater is the bias removal, but also the greater is the number of participants who must be dropped from the analysis because they cannot be matched.

The researcher is also confronted with another decision: how to form matches when more than one participant from the comparison group is within the given caliper distance from a participant in the treatment group. Two choices are most common: greedy and optimal matching. In greedy matching, a treatment group participant is chosen at random and is matched with the participant from the comparison group who has the nearest score on the covariate. If no match can be found within the caliper distance, the treatment group participant is dropped from the analysis. Another treatment group participant is then selected at random and matched, and so on. Obviously, who is matched with whom depends on the order in which the participants in the treatment condition are randomly chosen to be matched. It is possible that better matches could be found if participants were matched in a different order. Optimal matching avoids this problem and is to be recommended. With optimal matching, matches are created so that the overall distances between matched pairs are minimized across all the participants.

Researchers can also choose between matching with and without replacement. When matching without replacement, a comparison participant is removed from the matching pool once that participant has been matched. When matching with

replacement, a comparison participant can be used in more than one match. Compared to matching without replacement, matching with replacement can remove more bias due to selection differences on the covariate but can be less efficient because it might not use all the data and requires more complex analyses.

Researchers must also choose between one-to-one matching and one-to-many matching (West et al., 2014). With one-to-one matching, each participant in the treatment group is matched to a single participant from the comparison group. With one-to-many matching, a participant from the treatment group can be matched to more than one member of the comparison group, or vice versa (for an example, see Wu, West, & Hughes, 2008). One-to-one matching removes more bias than one-to-many matching because the matches tend to be closer. But one-to-many matching with replacement can be more efficient than one-to-one matching because the data are used more times, though power in one-to-many matching tends to asymptote when "many" gets to be five or six (West & Thoemmes, 2008). In what is called full matching, each treatment participant can be matched to more than one comparison participant, and vice versa. Full matching is often optimal (Rosenbaum, 2017).

Other matching options are also available, including methods to control for general dispositions that cannot be measured directly (Rosenbaum, 2013, 2015b, 2017). Such controls are obtained by "differential comparisons that overcompensate for observed behaviors in an effort to adequately compensate for an underlying disposition" (Rosenbaum, 2017, pp. 256–257). For example, consider a study of the effects of smoking that compares smokers to nonsmokers. Evidence suggests that smokers, more than nonsmokers, also tend to be disposed to engage in other risky behaviors, such as using hard drugs, which might lead to negative health outcomes. Being unable to directly measure the disposition for risky behavior, Rosenbaum (2017) suggested overcompensating for the dispositional difference by comparing smokers who never tried hard drugs to nonsmokers who had tried hard drugs.

Matching is a special case of blocking (also called stratification or subclassification—see Section 4.6.4). With blocking, participants from both treatment groups are placed into blocks based on their covariate scores, so that those within each block have similar scores on the covariate. The treatment effect is then estimated by comparing, within each block, the outcome scores of the participants from the two treatment groups and averaging the results across blocks after weighting the results by the sample sizes of the blocks. Matching attempts to equate participants perfectly on their baseline scores. Because participants within a block are usually not perfectly matched on their pretest scores, blocking does not perfectly equate participants on the pretest scores. As a result, blocking does not generally take account of selection differences as well as does matching. Nonetheless, blocking can do an admirable job of removing the effects of selection differences. The more blocks, the better is the control for selection differences. But Cochran (1968b; also see Cochran, 1965; Cochran & Rubin, 1973; Rosenbaum & Rubin, 1984) showed that placing participants into as few as five blocks removed 90% of the bias due to selection differences on the pretest scores. To increase power and

improve the quality of the matching, extra covariates can be added to the analysis as statistical controls after participants have been blocked.

When using multiple baseline measures, matching or blocking can be performed on each baseline measure individually, although matching or blocking on each baseline measure individually gets more complex as the number of baseline measures increases. As they increase, it becomes more difficult to find adequate caliper matches on all the variables. The alternative is to create a composite score of all the baseline measures and match or block using that composite. Several options are available for the composite. A traditional choice is the Mahalanobis distance metric, which measures the multidimensional distance of one participant's baseline scores to another participant's scores, taking account of the covariances among the measures. Currently, the most common method for creating a composite of baseline measures is to use **propensity scores** (see Section 7.6).

Because matching and blocking are doing the same thing as ANCOVA (i.e., equating participants on observed baseline measures), successful matching or blocking will produce much the same estimates of treatment effects as in ANCOVA, assuming the proper model is fit to the data using ANCOVA. As a result, whatever biases are present in ANCOVA (assuming a proper model is fit) will also be present with matching or blocking. In particular, matching and blocking (like ANCOVA) only copes with selection differences on the baseline measures that are included in the analysis. If other selection differences are present, matching and blocking (like ANCOVA) will be biased. That is, matching and blocking are no more likely to avoid hidden bias than is ANCOVA (see Section 7.4.1); there is no guarantee that hidden bias will be avoided. Also, like ANCOVA, matching and blocking techniques are biased by measurement error in the baseline variables. To avoid bias due to measurement error, the Porter correction for measurement error could be applied to a single baseline measure before participants are matched or blocked (Porter, 1967; see Section 7.4.2). That latent variable SEM (see Section 7.4.2) can take account of measurement error in multiple covariates is an advantage of the ANCOVA approach compared to matching or blocking. But again, such adjustments do not guarantee that the results will avoid hidden bias.

Compared to ANCOVA, one advantage of matching or blocking is that they do not require properly fitting a regression model to the data. Interactions and curvilinearity are taken into account automatically with matching and blocking. That matching or blocking might result in participants being dropped from the analysis may also be advantageous because it keeps the analysis from extrapolating beyond where data are present. ANCOVA can be applied even when the treatment groups have little or no overlap on the baseline measures, but such analyses might not result in credible treatment effect estimates. Blocking and matching avoid this problem because they can be implemented only when there is substantial overlap between the treatment groups on baseline measures. Ho, Imai, King, and Stuart (2007) provide an example where extrapolation with ANCOVA severely biased the estimate of a treatment effect. On the flip side, because participants might have to be dropped from the analysis with matching or

blocking, it might be difficult to know the population to which the treatment effect estimates apply. For example, the treatment effect estimate might not be either the ATE or the ATT (see Section 7.4).

Another significant advantage of matching or blocking compared to ANCOVA is that matching and blocking allows the data analysis to be specified without the use of the outcome scores, thereby eliminating fishing among a variety of analyses to obtain desired results (Ho et al., 2007; Rubin, 2007). That is, when applied to data, the ANCOVA model must include the outcome scores. Hence, in choosing which covariates, interaction terms, and polynomial terms to include in the ANCOVA model, the researcher has access to the estimate of the treatment effect. As a result, the researcher could, consciously or unconsciously, favor models that produce the more desirable treatment effect estimates. In contrast, matches and blocks can be established without reference to the outcome scores and hence without access to the estimate of the treatment effect. So with blocking and matching there is less opportunity, either consciously or unconsciously, to select the analysis model that provides the most desirable estimate of the treatment effect.

7.6 PROPENSITY SCORES

A propensity score is the probability that a participant is assigned to the treatment condition (rather than the comparison condition) based on the participant's scores on a given set of covariates (Austin, 2011; Rosenbaum & Rubin, 1983; Thoemmes & Kim, 2011). This set of covariates defines the propensity score. It is worth emphasizing the sense of the prior sentence: Propensity scores are defined for a given set of covariates. A different set of covariates defines different propensity scores. A propensity score of .75 means a participant has a 75% chance of being assigned to the treatment group, for the given set of covariates. A propensity score of .25 means a participant has a 25% chance of being assigned to the treatment group, for the given set of covariates.

The propensity score is a balancing score; that is, the distribution of the given set of covariates is the same in the two treatment groups, given the propensity scores (Rosenbaum & Rubin, 1983). As a result, adjusting for selection differences on the propensity score, by matching participants in the two groups on the propensity scores, is equivalent to adjusting for selection differences by matching participants on all the covariates that define the propensity score. It is important to understand the implications of the prior sentence. Rather than having to adjust for each covariate (in the set of covariates that define the propensity scores) individually, it is sufficient to adjust only on the propensity scores. In essence, propensity scores summarize the influence that all the covariates have on the selection of units into treatments. Thus, adjusting on propensity scores adjusts for all the covariates that define the propensity scores. This is one of the primary appeals of propensity scores. Instead of matching on each covariate in an entire set of covariates, the researcher needs to match only on the single variable of the

propensity scores (and, as noted, it can be difficult to achieve adequate matches when matching is performed on numerous covariates individually). Alternatively, instead of including an entire set of covariates into an ANCOVA model, the researcher needs to include only the propensity scores. Therefore, in ANCOVA, controlling for one variable (i.e., the propensity scores) substitutes for controlling for many variables (i.e., the covariates that define the propensity scores). If only a single baseline measure is available, however, using propensity scores offers little advantage.

If controlling for a given set of covariates is sufficient to remove the effects of all selection differences, then controlling for the propensity scores that are defined by the given set of covariates is sufficient to remove the effects of all selection differences on the outcome. In this case, the covariates produce selection differences that are ignorable (see Section 7.4.1). (Technically, I should say the assignment mechanism, rather than selection differences, is ignorable, but talking about selection differences is sometimes more intuitively clear than talking about assignment mechanisms.) The converse also holds. If controlling for the given set of covariates is not sufficient to remove the effects of all selection differences, then controlling for the propensity scores is not sufficient to remove the effects of all selection differences. Such selection differences are nonignorable. Hidden bias would remain.

With those caveats in mind, the researcher implements a propensity scores analysis in three steps: estimating the propensity scores, checking balance on the covariates and propensity scores, and estimating the treatment effect using the estimated propensity scores. If necessary, the first two steps are performed iteratively until balance is achieved (Austin, 2011).

7.6.1 Estimating Propensity Scores

The true propensity scores are unobserved in nonequivalent group designs. Hence, the propensity scores for a given set of covariates must be estimated. The most common way to estimate propensity scores is to use logistic regression to predict treatment assignment (entered as an indicator variable: 1 if the participant is in the treatment group and 0 if the participant is in the comparison group) based on the given covariates that are used to define the propensity scores. The predictors in the logistic model may include polynomial terms as well as interactions among the covariates. In this way, the polynomial terms and the interactions are included in the set of covariates that define the propensity scores. Sometimes more complex methods (such as boosting and random forests) have been used to deal with nonlinearities between covariates and propensity scores instead of logistic regression (Austin, 2011).

If the propensity scores are to account for all selection differences, the set of covariates needs to include all variables that account for selection and are related to the outcome measure (Austin, 2011; May, 2012; Stuart, 2010). If the propensity scores cannot be well estimated based on the necessary set of covariates (especially if the propensity scores only weakly predict selection), a propensity score analysis might increase, rather

than decrease, bias due to selection differences (Shadish, Clark, & Steiner, 2008; Shadish & Cook, 2009).

7.6.2 Checking Balance

After propensity scores have been estimated, the researcher matches or blocks (see Section 7.5) participants in the treatment conditions on the propensity scores and checks to make sure the treatment groups are balanced on the covariates and propensity scores. Balance means the joint distribution of the covariates is the same across the treatment groups for participants matched on the propensity scores. Balance would be guaranteed if the true propensity scores were known and matching were perfect. As noted, however, the true propensity scores are unknown and must be estimated. Because the propensity scores must be estimated and because matching on the propensity scores may not be perfect, balance might not be well obtained. So, the next step in a propensity score analysis is checking on the estimation of the propensity scores by checking on the balance of the participants in the two treatment conditions.

Balance is assessed by matching or stratifying the participants in the treatment groups on the propensity scores. Then the distributions of the covariates (and the propensity scores) are compared across the matched or stratified groups. Technically, balance should be assessed on the covariates (and propensity scores) jointly—that is, by examining their multivariate distributions. In practice, however, balance is usually assessed on each covariate (and the propensity scores) individually or, at most, by examining bivariate distributions. Included in checks of balance are comparisons between the treatment groups of entire frequency distributions as well as of the means and variances of the covariates and propensity scores. If the propensity scores are estimated using polynomial terms or interactions of the covariates, balance is also checked on the polynomial and interactions terms. If the participants are stratified into blocks rather than matched on the propensity scores, researchers can assess balance by performing a two-way ANOVA where treatment groups and blocks are the factors and covariates are the dependent variables. The presence of a main effect due to treatment group or an interaction of treatment group and block indicates a lack of balance. Whether balance has been achieved is best determined using measures such as Cohen's d (which is a standardized mean difference and is derived as the group-mean difference divided by the within-group standard deviation) rather than levels of statistical significance of differences because the latter is sensitive to sample size (Stuart, 2010). A rule of thumb is that Cohen's d should not be greater than 0.25. But some researchers use the more stringent criteria of a Cohen's d not greater than 0.1 (Austin, 2011). Balance is also usually assessed by comparing the ratio of the variances of the covariates in the two treatment groups. A rule of thumb is that such ratios should be between the values of 4/5 and 5/4 (Rubin, 2007). A researcher could also compare the entire distributions of the baseline variables for the two treatment conditions using quantile–quantile (QQ) plots (Ho et al., 2007).

If balance is inadequate, the researcher reestimates the propensity scores by including additional covariates or by adding polynomial and interaction terms of the covariates already in the estimation model. Indeed, even if balance is pretty good after a first attempt, researchers should purposely sift through numerous propensity score models and matching/blocking methods to produce the best balance possible (Ho et al., 2007; Rubin, 2007). To achieve the best balance possible, a researcher may need to include covariates in the creation of the propensity scores without regard to the statistical significance of the covariates in predicting the propensity scores. However, just as with ANCOVA, researchers should only include covariates that have not been affected by the treatment and are reasonably related to the outcome (Ho et al., 2007).

In assessing balance, the researcher first assesses the degree of overlap between the treatment groups on the propensity scores. Overlap is assessed by comparing the frequency distributions of the propensity scores between the treatment groups before the participants are matched or blocked on the propensity scores. Such a comparison can be drawn by plotting the frequency distributions of the propensity scores in the treatment groups, one on top of the other. The range of propensity scores over which there is overlap is called the **region of common support.** Without sufficient overlap (i.e., without an adequate region of common support), participants from the treatment groups cannot be matched or blocked on the propensity scores. Hence, balance cannot be assessed, and the statistical analysis cannot proceed. In any case, estimates of the treatment effect apply only to those participants included in the analysis (which might not equal either the ATE or ATT effects; see Section 7.4).

7.6.3 Estimating the Treatment Effect

Once balance has been obtained, the effect of the treatment is estimated using one of several methods (Austin, 2011; Stuart, 2010). The propensity scores (or usually the linear propensity scores rather than the estimated probabilities [Rubin, 2007]) can be used in a matching or blocking analysis (see Section 7.5). Or the propensity scores can be used as a covariate in an ANCOVA (see Section 7.4). In ANCOVA, the estimated propensity scores would be entered like any other covariate and could be included along with polynomial and interactions terms in the estimated propensity scores. Rubin (2007) and Ho et al. (2007) explain the several benefits of matching versus ANCOVA.

The treatment effect can also be estimated by the inverse weighting of the estimated propensity scores (Steiner & Cook, 2015). With inverse weighting, the outcome scores are divided by the estimates of the propensity scores. In this way, the outcome scores of underrepresented participants (i.e., those with small propensity scores) are upweighted, while the outcomes scores of overrepresented participants (i.e., those with large propensity scores) are downweighted. A disadvantage of inverse weighted is that it can produce unreliable results when some propensity scores are close to either zero or one.

Other covariates, besides the propensity scores, can be added to the analysis (Rosenbaum, 2017; Rubin, 2004; Schafer & Kang, 2008; Stuart, 2010) either when

matching/blocking or using ANCOVA. Indeed, research shows that the best method can be to combine matching/blocking with regression adjustments (Schafer & Kang, 2008; Stuart & Rubin, 2007). Matching and blocking may have residual bias because of the lack of perfect matches on the propensity scores. One purpose of adding more covariates is to reduce that bias; another purpose is to increase power and precision. A propensity score analysis is a model of the selection of participants into treatment groups and does not necessarily well model the causal determinants of the outcome measure. Adding covariates that account for the outcome measure and not just for selection can increase the power and precision of the analysis. In addition, adding covariates can make the analysis doubly robust. Remember: to remove bias due to selection differences, the covariates must model either the outcome scores, the selection differences, or both (see Section 7.4.1). Propensity scores model selection differences. Additional covariates could be included in the statistical analysis to model the outcome scores. In this way, the estimates of treatment effects would be unbiased if either model is adequate. Another purpose in adding covariates to the analysis is to improve design sensitivity (Rosenbaum, 2017; see also Section 7.9). Matching on covariates that are highly predictive of outcomes can reduce the heterogeneity of matched pair outcomes, which increases the robustness of the design to hidden bias.

7.6.4 Bias

Propensity scores are not a cure-all for the biasing effects of selection differences. As noted earlier, the primary advantage of propensity scores is that they allow researchers to control for all the covariates that go into the construction of the propensity scores using a single variable (the propensity score) rather than each of the covariates individually. But using propensity scores removes no more nor less bias than using all the covariates individually. If controlling for the individual covariates is not sufficient to remove bias due to selection differences, propensity scores are not sufficient to remove bias due to selection differences. The advice usually given (Stuart, 2010) is to err on the side of including more rather than fewer covariates when estimating propensity scores, so as to control for as much bias as possible with the available covariates.

Also recall that measurement error in the covariates leads to bias in that the effects of selection differences are not completely removed when the covariates are fallible. The same bias exists with propensity scores. If propensity scores are estimated based on fallible covariates, the propensity scores will not remove all bias due to selection differences on those covariates (Cook & Steiner, 2010; Cook, Steiner, & Pohl, 2009; Steiner, Cook, & Shadish 2011). According to Steiner et al. (2011), every 10% reduction in the reliability of a covariate reduces by 10% the proportion of bias reduction for that covariate. Multiple covariates can compensate to some extent for unreliability in each individual covariate, but the degree of compensation depends on the correlations among the covariates.

There is some disagreement, when using propensity scores, about which method of estimation of the treatment effect (whether ANCOVA, matching, blocking, or inverse weighting) best removes bias due to selection differences (Cook et al., 2009; Shadish et al., 2008; Steiner & Cook, 2015). In addition, Steiner and Cook (2015, pp. 255–256) note: "Despite the theoretical advantage of [propensity scores] with regard to design and analytic issues, it is not clear whether they actually perform better in practice than standard regression methods (i.e., regression analyses with originally observed covariates but without any [propensity score] adjustment)." As also noted in Section 7.12.3, the method of estimation appears to be not nearly as important for removing bias due to selection differences as are the covariates included in the analysis (Cook et al., 2009; Pohl, Steiner, Eisermann, Soellner, & Cook, 2009; Shadish et al., 2008). That is, the quality of the covariates drives the ability of propensity scores to remove bias due to selection differences more than does the choice among methods of estimation. As West et al. (2014, p. 915) state: "Selecting a comprehensive set of covariates to be measured at baseline is the *most* critical issue in propensity score analysis." In most cases, the most important measures to include in the analysis are pretests that are operationally identical to the posttests and covariates that account for selection and are highly correlated with outcomes (Cook & Steiner, 2010, Cook et al., 2009; Steiner, Cook, Shadish, & Clark, 2010; Wong, Wing, Steiner, Wong, & Cook, 2012).

Of course, other complications can arise and introduce additional complexity into the analysis. Cham and West (2016) present various approaches to propensity score analyses when data are missing from both covariates and outcome measures, for example.

Estimating treatment effects using propensity scores is likely the most commonly used method of data analysis for the nonequivalent group design. Nonetheless, as with all methods, it is not without limitations and critics (e.g., King & Nielsen, 2016; Peikes, Moreno, & Orzol, 2008).

7.7 INSTRUMENTAL VARIABLES

Another way to remove the effects of selection differences is with instrumental variables (Angrist et al., 1996; Angrist & Krueger, 2001; Bollen, 2012; Sovey & Green, 2011). Consider the regression of Y onto the indicator variable T that represents treatment assignment using the simple ANOVA model:

$$Y_i = \alpha_Y + (\beta_T T_i) + u_i \tag{7.6}$$

As an estimate of the treatment effect, the estimate of the β_T, is susceptible to bias due to selection differences. One way to conceptualize the problem is as an omitted variable (Reichardt & Gollob, 1986). Suppose W is a pretreatment variable that completely

accounts for selection differences, so that if the participants in the two groups were matched on that variable, the posttest differences between those matched pairs would provide an unbiased estimate of the treatment effect. That is, under these conditions, selection differences would be ignorable given *W,* so the following model (assuming *W* is linearly related to *Y*) would produce an unbiased estimate of the treatment effect:

$$Y_i = \alpha_Y + (\beta_T T_i) + (\beta_W W_i) + v_i \tag{7.7}$$

From this perspective, Equation 7.6 produces a biased estimate of the treatment effect because *W* has been omitted from the model and *W* is correlated with *T* because the groups differ on initial characteristics. Now suppose *W* is unavailable so that Equation 7.7 cannot be fit. You can still obtain an unbiased estimate of the treatment effect if you have an instrumental variable, *Z,* to replace *T*. To be an adequate instrumental variable, *Z* would have to substantially influence the treatment assignment, *T* (which is called the **relevance assumption**). The instrument (*Z*) would also have to influence the outcome variable but only through *T*. This is called the exclusion restriction because the instrument is excluded from Equation 7.7, which would produce an unbiased estimate of the treatment effect were *W* available. In addition, the instrument must not be correlated with the omitted variables *W* or with the error term v_i. This is called the **independence assumption.** Finally, the effect of the instrument cannot be to push some participants into treatment while it also pushes others out, which would mean there are no defiers (see Section 4.7.4). This is called the monotonicity assumption.

If you think in causal terms, you could describe the situation as in Figure 7.4, which is a path diagram among the variables (May, 2012). In Figure 7.4, an arrow means that one variable causes another and a double-headed curved arrow means that two variables are correlated. Figure 7.4 indicates that *Z* causes *T* (with regression coefficient β_Z), which in turn causes *Y* (with regression coefficient β_T). But there is no direct effect of *Z* on *Y* (i.e., no direct arrow from *Z* to *Y*). In other words, the treatment (represented by *T*) completely mediates the effect of *Z* on *Y*. Figure 7.4 also indicates that *W,* which

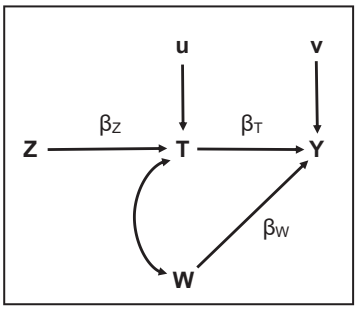


FIGURE 7.4. A path diagram revealing the relationships among the instrumental variable (*Z*), the treatment variable (*T*), the posttest (*Y*), and a covariate (*W*). Arrows represent causal effects. The curved, double-headed arrow indicates a correlation.

represents unmeasured selection differences, is correlated with T and causes Y (with regression coefficient β_W), but is not correlated with Z. Also note the implication that if W were available, Y could be regressed onto both T and W to obtain an unbiased estimate of the effect of T on Y, which is the treatment effect.

An instrumental variable (IV) approach operates in the following fashion. Under the conditions shown in Figure 7.4, regressing Y onto Z (called the reduced-form equation) produces a regression coefficient that is equal to the product of the estimates of β_Z and β_T. Then regressing T onto Z (called the first-stage regression) produces the estimate of the regression coefficient β_Z. Dividing the product of the estimates of β_Z and β_T by the estimate of β_Z gives the estimate of β_T, which is the effect of the treatment in the form of the local average treatment effect (LATE) estimate where the LATE estimate is the average effect of the treatment on those for whom the instrument influences the treatment status—who are said to be the compliers (Angrist & Pischke, 2009). If you think of the instrument as providing an incentive or nudge to participate in the treatment, the compliers are those who participate in the treatment but only if given the incentive. For this reason, the value of the LATE depends on the instrument.

The first-stage and reduced-form equations can be used to check the assumptions of the analysis (Angrist & Pischke, 2015). The relevance assumption demands that the regression coefficient in the first-stage regression be substantially nonzero. A weak relationship suggests a weak instrumental variable (see below). The exclusion restriction can only be tested indirectly. If the first-stage regression coefficient is zero, the reduced-form coefficient should be zero as well. But a reduced-form coefficient that is nonzero accompanied by a first-stage coefficient that is zero suggests a violation of the exclusion restriction. The independence assumption can be assessed indirectly by looking for relationships between the instrumental variable and available covariates.

In practice, the coefficient β_T is usually estimated using two-stage least squares (2SLS), which operates by simultaneously fitting slightly different regressions than noted above but employs the same logic. The first regression model estimates T given Z:

$$T_i = \alpha_T + (\beta_Z Z_i) + u_i \tag{7.8}$$

Estimated values of T ($\hat{T}$) are then calculated as

$$\hat{T}_i = a_T + (B_Z Z_i) \tag{7.9}$$

where a_T and B_Z are the estimated values of α_T and β_Z from Equation 7.8. In the second regression model, Y is regressed onto $\hat{T}$:

$$Y = \alpha_Y + (\beta_T \hat{T}_i) + e_i \tag{7.10}$$

where the estimate of β_T in this last regression is the estimate of the LATE and where the 2SLS procedure (by fitting the two equations together) makes an adjustment to the

estimated standard error of the estimate of β_T to take account of having performed the first regression. (If you fit the two regressions separately, you get an incorrect estimated standard error; you should use the 2SLS simultaneous procedure.) In essence, at the second stage, you are estimating the effect on *Y* of that part of *T* that is not correlated with the selection differences represented by *W*. It is also possible to obtain the same estimate of the treatment effect using a SEM approach (Murnane & Willett, 2011). Covariates can be added as control variables, but the same covariates must be added to both Equations 7.8 and 7.10.

For an example of the use of instrumental variables, suppose you wanted to estimate the effects of smoking on lung cancer (Leigh & Schembri, 2004). The problem is selection differences: those who smoke differ in innumerable ways from those who do not smoke, and any relationship between smoking (*T*) and lung cancer (*Y*) could be due to the influence of these selection differences (*W*) rather than to the effect of smoking. What you need is an instrument (*Z*), which means a variable (a) correlated with smoking and thus correlated with lung cancer but without having a direct effect on lung cancer and (b) not correlated with selection differences. The local tax rate on cigarettes could be related to smoking because higher taxes might reduce the amount spent on cigarettes. But the tax rate is presumed not to affect health outcomes directly nor to be correlated with selection differences.

Instrumental variable approaches have two drawbacks. First, they are large-sample methods. The instrumental variable approach produces only consistent—not unbiased—estimates of the treatment effect. Results are close to being unbiased only in large—perhaps very large—samples (Angrist & Pischke, 2009).

Second, it may be difficult to find a variable that qualifies as a strong instrument. That is, it may be difficult to find a variable *Z* substantially correlated with *T,* yet correlated with *Y* only via its correlation with *T* and uncorrelated with selection differences *W*. Consider the smoking example again. Questions can be raised about whether tobacco taxes are an adequate instrument. It is possible that tobacco taxes are correlated with outcomes independently of their relation to smoking. For example, perhaps communities that impose higher tobacco taxes are also those that are more health conscious in general. In that case, tobacco taxes would be related to health, even holding smoking constant. In this case, using tobacco taxes as an instrument would not produce consistent estimates of the treatment effect. In practice, instruments have most often been policies or circumstances that are gatekeepers to acquiring the treatment (such as month of birth for enrollment in school, proximity to treatment services, or differences in admission policies in different locales). Such instruments, however, are not available in many practical applications. As Rosenbaum (2017, p. 278) notes, "Instruments that meet [the] requirements are hard to find, and it is difficult to marshal evidence demonstrating that one has been found." As Schafer and Kang (2008, p. 306) observe about psychological research, "The presence of an instrument tends to be the exception rather than the rule." Instruments are more commonly used in research in sociology and economics than in psychology.

In addition, even though tobacco taxes are likely correlated with smoking, they are unlikely to be substantially correlated. That is, the instrument, *Z*, is only weakly correlated with the treatment assignment, *T*. Because smoking is addictive, it may not be sensitive to changes in the price of cigarettes. An instrument that is only weakly related to treatment assignment *T* is called a weak instrument. In logic, weak instruments are adequate for the technique to work. But, in practice, weak instruments tend to produce unstable estimates of treatment effects that can be substantially biased. Moreover, weak instruments produce designs that are highly sensitive to hidden bias (Rosenbaum, 2017; see Section 7.9).

7.8 SELECTION MODELS

The selection model approach is credited to Heckman (1979; Heckman & Robb, 1986). Heckman's method was originally developed for dealing with sample selection wherein a researcher has a nonrandom, truncated sample from a population and wishes to estimate regression parameters for the whole population. But the method was subsequently applied to deal with selection differences in the nonequivalent group design.

The classic specification of the approach involves two equations: the outcome and the selection equations (Briggs, 2004). The outcome equation is a model of the outcome of interest:

$$Y_i = \alpha_Y + (\beta_T T_i) + (\beta_X X_i) + \ldots + \varepsilon_i \tag{7.11}$$

where

Y_i is the outcome variable;
T_i is the treatment variable, which is equal to 1 if the *i*th participant is in the treatment condition and 0 if the *i*th participant is in the comparison condition;
X_i is an observed covariate that predicts the outcome and the ellipses indicate there can be more than one; and
ε_i is the residual or error term, which is assumed to be independent of the *X*'s.

If T_i were unrelated to the error term given the covariate(s), the ordinary least squares estimate of β_T would be an unbiased estimate of the treatment effect. The problem is that there might be unmeasured variables that are correlated with treatment assignment (*T*) and that influence the outcome (*Y*). This would make *T* correlated with the error term (ε), even after controlling for the measured covariates, which would produce a bias in the treatment effect estimate (Reichardt & Gollob, 1986). Such a bias is due to selection differences arising from variables omitted from the model—that is, hidden bias. If the expected value of ε given both *T* and the *X*'s were known, this expected value (γ) could be included in the model to obtain an unbiased estimate

of the treatment effect (Briggs, 2004). In other words, the expanded outcome model would be

$$Y_i = \alpha_Y + (\beta_T T_i) + (\beta_X X_i) + \ldots + (\beta_\gamma \gamma_i) + v_i \tag{7.12}$$

which would produce an estimate of β_T that was an unbiased estimate of the treatment effect. Of course, the variable γ is not known, but it can be estimated using a selection equation, which is a model for the selection of participants into the treatment:

$$\begin{aligned} T_i &= 1 \text{ if } \alpha_T + (\beta_Z Z_i) + \ldots + u_i > 0 \\ &= 0 \text{ otherwise} \end{aligned} \tag{7.13}$$

where

Z_i is an observed covariate that predicts selection and the ellipses indicate there can be more than one; and
u_i is the residual or error term, which is assumed independent of the Z's.

The variable γ in Equation 7.12 is estimated as an inverse Mills ratio (which is equal to a probability density function divided by a cumulative distribution function) based on the value of Z or Z's in Equation 7.13. Under the additional (strong) assumption that the errors ε and u are bivariate normally distributed, the value of γ can be estimated for each participant by fitting Equation 7.13 using probit analysis. Then Equation 7.12 can be fit using the estimated γ to produce a consistent (though biased) estimate of the treatment effect β_T (Briggs, 2004; Stolzenberg & Relles, 1997). Equation 7.12 is fit using generalized least squares because the v's are heteroskedastic. Alternatively, the whole procedure can be fit in a single step, rather than in two steps, using maximum likelihood estimation. Other specifications of the models are also possible (Barnow, Cain, & Goldberger, 1980). In addition, the treatment effect estimate can be improved by including in the set of Z variables in the selection Equation 7.13 some variables that are highly correlated with selection into treatment conditions but are not any of the X's that appear in Equations 7.11 or 7.12 (the exclusion restriction). Such covariates serve as (strong) instrumental variables.

Regardless of how the equations are specified, the problem is that strong assumptions are required in deriving the estimate of the treatment effect and at least some of these assumptions are not directly testable and may well not be correct. Unfortunately, the estimate of the treatment effect β_T has been shown to be highly sensitive to violations of the assumptions and to the choice of covariates included in the equations (Briggs, 2004; Stolzenberg & Relles, 1997). Indeed, when the assumptions are violated, the results from selection modeling can produce worse estimates than estimating treatment effects from Model 7.11 rather than Model 7.12. In addition, there can be a large degree of multicollinearity in Model 7.12 because T and γ are often highly correlated,

which reduces power and precision. Although selection models have been very popular, it appears that they have been largely superseded by propensity score analyses and instrumental variables approaches in addressing selection bias in nonequivalent group designs, as well as by other quasi-experimental designs (Cook & Wong, 2008a)—though see DeMaris (2014).

7.9 SENSITIVITY ANALYSES AND TESTS OF IGNORABILITY

There are no direct empirical tests of whether hidden bias is present and how much it might affect the estimates of treatment effects (see Section 7.4.1). The best researchers can do is assess the effects of hidden bias indirectly. There are multiple approaches to assessing the effects of hidden bias (Liu, Kuramoto, & Stuart, 2013). In the following, I describe three methods that fall into two categories.

The methods in the first category assess the degree to which estimates of a treatment effect are sensitive to hidden bias. Two methods in this category (called **sensitivity analyses**) are presented below. Depending on the results of the sensitivity analyses, a design is said to be either sensitive or insensitive to various sizes of violations of the assumption of ignorability. Designs can be made more resistant to violations in at least four ways: (1) make the treatment effect large; (2) reduce the heterogeneity of matched pair differences in outcomes; (3) match each treatment participant to more than one comparison participant; and (4) assign treatments to clusters of participants (Rosenbaum, 2017). Note, however, that design sensitivity is not revealed by either the size of a treatment effect estimate or either the *p*-value or confidence interval for a treatment effect estimate—you need to perform the calculations given below.

The second category of methods consists of tests for the presence of hidden bias (or conversely, tests for ignorability) that are conducted using additional comparisons. Consider each of these three methods in turn (Rosenbaum, 1986, 1987, 2002, 2005b, 2010). Also see Manski and Nagin (1998) for setting bounds on effect size estimates under different assumptions about treatment selection.

7.9.1 Sensitivity Analysis Type I

The first type of sensitivity test assesses how much the estimate of a treatment effect would change by making assumptions about the nature of omitted variables. In other words, this type of sensitivity analysis asks how much the results would be biased if covariates that should be included in the analysis were instead omitted from the analysis, assuming those omitted covariates had certain characteristics.

Such a sensitivity analysis can be conducted in the following fashion (Rosenbaum, 1986, 2002; West & Thoemmes, 2008). Let variable U be an aggregate of all the omitted covariates so that if U were included in the analysis, the estimate of the treatment effect would be unbiased. Then assume that U (1) is as highly correlated with the outcome as

is the most highly correlated observed covariate and (2) has a standardized mean difference between the treatment groups that is as large as the largest standardized mean difference of any observed covariate. Then calculate how much the confidence interval and the statistical significance test of the estimate of the treatment effect would change if the *U* variable had been included in the analysis. If the changes are not enough to alter the substantive conclusions, then the results of the analysis are said to be robust to omitted covariates.

For example, assume that the largest correlation between an observed covariate and the outcome variable is .5. Also assume that the observed covariates have a standardized mean difference between the treatment and control groups that is no greater than 1.0. Then calculate the confidence interval for the treatment effect assuming *U* is correlated .5 with the outcome and has a standardized mean difference of 1.0 and was included in the analysis. If zero is not contained in that confidence interval, the statistical significance of the estimate of the treatment effect is robust to bias due to omitted variables of this nature.

7.9.2 Sensitivity Analysis Type II

The second type of sensitivity analysis determines how large a bias due to an omitted covariate would need to be for the estimate of the treatment effect to no longer be statistically significant. Such a sensitivity analysis is conducted in the following fashion (May, 2012; Rosenbaum, 2005b).

Assuming the results of the test of an estimate of the treatment effect were statistically significant at the .05 level, let gamma be the odds ratio of receiving the treatment condition rather than the comparison condition. For example, if gamma were equal to 2, a participant would be twice as likely as another (in terms of odds) to receive the treatment condition rather than the comparison condition. In a simple randomized experiment, gamma would be 1. Under the assumption of strong ignorability, gamma would also be 1 after matching on observed covariates. But if hidden bias is present, gamma could be greater than 1. Now calculate the confidence interval and the obtained probability (*p*) value for different values of gamma. The critical gamma is the value of gamma where the endpoint of the confidence interval for the treatment effect equals zero and the *p*-value is correspondingly equal to .05. A gamma larger than the critical gamma would mean the results would no longer be statistically significant at the .05 level. Then if the value of this critical gamma is larger than seems likely to arise for omitted variables (given the observed covariates in the analysis), the researcher can be plausibly assured that the treatment effect would be statistically significant, even if omitted covariates were included in the analysis. In other words, this form of sensitivity analysis assesses the degree of hidden bias as represented by the gamma that could arise before the confidence interval for the effect size would contain zero, or, equivalently, the *p*-value of a statistical significance test would be greater than .05. May (2012, p. 506) provides an example:

> If a sensitivity analysis yielded a [critical gamma] of 6 as in the example, then one would conclude that to dismiss the estimate of treatment effect as being caused by selection bias, one or more unmeasured confounds would need to exist that would make the subjects in the treatment group at least 6 times more likely to experience the treatment than those in the comparison group. Given that odds ratios of 6 are quite uncommon in social science research, it is reasonable to conclude that the unobserved selection bias would need to be gigantic in order to dismiss the estimated treatment effect.

To gauge what gammas are likely to arise, a researcher might consider the gammas that arise for observed covariates.

7.9.3 The Problems with Sensitivity Analyses

Sensitivity analyses can place plausible limits on the size of the effects of hidden bias, but they cannot rule out the possibility of hidden bias with certainty. Sensitivity analyses are fallible. For example, Rosenbaum (1991) reports a study in which the results seemed to be robust to the effects of hidden bias, but later evidence from a randomized experiment suggested that hidden bias was substantially larger than assumed in the sensitivity analysis (see also Steiner et al., 2008). Sensitivity analyses are fallible because the required assumptions about the maximum size of omitted selection biases may be incorrect.

7.9.4 Tests of Ignorability Using Added Comparisons

Ignorability cannot be tested directly. The best that can be done is to conduct a test that provides an opportunity for results to be contrary to ignorability (Rosenbaum, 1984a, 1987). Then if results are indeed contrary to ignorability, hidden bias is assumed to be present. If, however, the empirical results are consistent with ignorability, it does not mean hidden bias is ignorable; rather, it merely indicates that a test of ignorability has succeeded. The logic of the test accords with Popper's principle of falsifiability whereby theories are never proven correct, they are simply not falsified after being subjected to severe tests. Passing the test makes unconfoundedness more plausible but not certain. Such tests are called **falsification tests.**

For example, consider a study of the effects of dropping out of school on academic achievement (Rosenbaum, 1986). If dropping out has an estimated effect that is negative as presumed, the effect should be related to the time at which students drop out. A student who drops out after 10th grade should perform worse than a student who drops out after the 11th grade. So, in addition to a global comparison of those who drop out to those who do not drop out, add a comparison of students who drop out at various grade levels. A similar test can be performed with the effects of smoking. In addition to comparing smokers to nonsmokers, also compare those who smoked (or have quit smoking) for different lengths of time and those who smoked different amounts. In both cases, the hypothesized effects should be present if either dropping out or smoking has effects. If the predicted results do not arise, however, then hidden biases might be present.

Both Rosenbaum (1984a) and Steiner and Cook (2015) provide additional examples of tests of ignorability. The first test uses a nonequivalent dependent variable, which is similar to the outcome measure of interest but is presumed not to be affected by the treatment. That means a null result should be found if an analysis to detect a treatment effect were applied to the data from the nonequivalent dependent variable. If a null result is obtained, on the one hand, ignorability remains plausible. On the other hand, ignorability is placed in doubt to the extent that an apparent treatment effect is found for the nonequivalent outcome measure. Such a spurious treatment effect is often called a placebo outcome, but I call it a **pseudo treatment effect.**

A second test requires two nonequivalent comparison groups that receive no treatment (Rosenbaum, 1987, 2017). The test of ignorability has been passed to the extent that the two comparison groups exhibit no outcome differences (no pseudo treatment effects because there are no differences in treatments between the two comparison groups) after controlling for observed covariates. To create two such comparison group contrasts, a researcher might compare those who were offered treatment but declined to accept it to those who were not offered the treatment and therefore did not receive it. **Dry-run analyses** (see Section 7.11.5) also provide means to assess ignorability.

7.10 OTHER THREATS TO INTERNAL VALIDITY BESIDES SELECTION DIFFERENCES

Selection differences are inherent in the nonequivalent group design. Because the performances of different groups of participants are being compared, selection differences are always present and should always be taken into account in the statistical analysis of data from the nonequivalent group design. It is for this reason that taking account of selection differences is the focus of the present chapter. Such a focus is generally shared by others when they discuss the nonequivalent group design. That is, because selection differences are a primary and ever-present threat to internal validity in a nonequivalent group design, the literature on the nonequivalent group design usually focuses on the means of coping with selection differences rather than the means of coping with other threats to internal validity. This chapter is no different.

Nonetheless, other threats to internal validity, besides selection differences, are also possible in the nonequivalent group design, just as they are in randomized experiments, and deserve mention. Differential attrition (also called selection by attrition) is a threat to internal validity in a randomized experiment (see Section 4.8) and can also occur in a nonequivalent group design. Those who attrit from the comparison group might be those least likely to improve (perhaps because they are least motivated and leave the study because they are discouraged by not receiving the treatment), whereas those who attrit from the treatment group might be those who would show the greatest improvement and feel they no longer need the treatment being offered. Or it could be the reverse. That is, those who attrit from the treatment group might be those least likely

to improve, leaving because they are discouraged since they have shown no improvement from the treatment, and so on. In either case, differential attrition simply adds to the nature of nonequivalence between the treatment groups and can be addressed using the methods for dealing with selection differences given in the present chapter. Missing data methods such as multiple imputation or full information maximum likelihood might also be helpful but, as noted, are not a cure for data that are not missing at random (see Section 4.8.2).

Noncompliance to treatment conditions can also be a threat to internal validity. Participants can fail to participate in the treatment to which they are initially assigned or to which they initially self-select. But, again, noncompliers (i.e., crossovers and no-shows) simply add to the nature of the nonequivalence between the treatment groups, and their effects can be addressed using the methods of addressing nonequivalence given above. That is, nonequivalence due to noncompliance just becomes another form of initial nonequivalence between treatment groups, and both forms of nonequivalence are addressed together using the methods described in this chapter. The analysis becomes an analysis of treatment as received. The analysis compares those who received the treatment to those who did not, regardless of the initial assignment. Such an analysis is not a recommended approach in a randomized experiment but is the default option in the nonequivalent group design.

Another potential threat to internal validity is differential history (also called local history or selection by history). Differential history arises when different external events affect the treatment groups. For example, either the treatment or comparison group might be subject to an additional intervention—perhaps because an administrator wishes to add another feature to the treatment or to compensate the comparison group for not having the treatment. Or perhaps a researcher interested in estimating the effects of a nutrition program finds that those eligible for the program are also eligible for food stamps, which, therefore, is confounded with the intended intervention. The best way to address differential history is to design the study so that the threat is minimized or avoided.

Instrumentation may also differ across the treatment conditions (called differential instrumentation or selection by instrumentation). By that I mean the measurement methods used to record observations are different in the two treatment conditions. Just as in a randomized experiment, if measurements are made by different observers in the different treatment conditions, the observers in the comparison condition may become more bored and make less careful observations than the observers in the treatment condition. Alternatively, perhaps the observers are aware of the research hypothesis being investigated and, as a result, come to have different expectations about how the results should turn out in the different treatment conditions. Then because of confirmation biases, the different expectations might lead the observers to find the results they expect to arise, even if such differences are not objectively present. Again, the best way to address differential instrumentation is to design the study so that the threat is minimized or avoided.

Differential testing (or selection by testing) would arise if one group were given more pretreatment measures than the other group. Again, the best way to address differential testing is to minimize or avoid the threat by the design of the study. Threats may also be due to differential maturation (or selection by maturation) and differential regression toward the mean (or selection by regression). These latter two threats, already mentioned in Section 7.3, are threats to validity for the change score analysis, which assumes that change over time is equal across the treatment groups in the absence of a treatment effect. Other analysis strategies do not assume equal change over time across the treatment groups and at least have the potential to take account of such threats to validity. For example, under conditions of ignorability, both ANCOVA and matching/blocking correct for differences between the treatment groups in maturation and regression toward the mean.

7.11 ALTERNATIVE NONEQUIVALENT GROUP DESIGNS

The prototypical nonequivalent group design that has been considered so far can be modified in a variety of ways. Consider a few of these variations.

7.11.1 Separate Pretest and Posttest Samples

In the prototypical nonequivalent group design, the same participants are assessed on the pretest and posttest measures. The participants measured at pretest need not, however, be the same as those measured at posttest. Instead, the participants measured at pretest and posttest could be independent samples. For example, pretreatment observations might be collected from samples of students at the start of a school year, with the posttreatment observations assessed on different samples of students at the end of the school year. The design is diagrammed as follows:

$$\begin{array}{lll} \text{NR: } O_1 & \mid \quad X & \quad O_2 \\ \text{NR: } O_1 & \mid & \quad O_2 \end{array}$$

where the vertical lines indicate that the pretest and posttest observations are collected on separate samples of participants. To make sense, operationally identical measures would have to be used at both pretest and posttest.

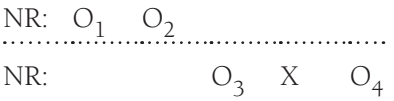
This design has the advantage of avoiding testing effects. Its drawback is that ANCOVA and matching procedures cannot be used. The only obvious analysis option is a DID method implemented with a two-by-two ANOVA (see Equation 7.3 in Section 7.3 and Angrist & Pischke, 2009). Thus, there would be no adjustments for initial selection differences except to assume that the effects of selection differences were identical

at pretest and posttest. In the absence of a treatment effect, the mean pretest differences between the treatment groups would be the same as the mean posttest differences.

7.11.2 Cohort Designs

The more similar the groups are to begin with, in a nonequivalent group design, the less room there is for bias due to selection differences. One way to make groups similar is to use cohorts (see Section 6.4). For example, a program could be introduced to all sixth graders in a school in, say, the academic year 2017–2018. In this case, the treatment group would be students who were in sixth grade during the academic year 2017–2018, while the comparison group could be students from the same school who had been in the sixth grade in the academic year 2016–2017. Comparisons of siblings provide other examples of cohort designs (Currie & Thomas, 1995; Lahey & D'Onofrio, 2010).

A cohort design can be diagrammed thusly:



The diagram shows that the pretests and posttests in the treatment groups are not collected at the same times. To return to the previous example, pretest and posttest for the sixth graders in the treatment group would be measured at the beginning and end of the 2017–2018 academic year, while the pretest and posttest for the comparison group of sixth graders from the same school would be measured at the beginning and end of the 2016–2017 academic year. Shadish et al. (2002) use a dotted line (as I do above) to separate the two cohorts rather than a dashed line as in the prototypical nonequivalent group design. As noted before, the dotted line is meant to emphasize that the cohorts are likely to be more similar than noncohort groups in the typical nonequivalent group designs. (Cook & Campbell [1979] used wavy lines instead of dotted lines for the same purpose.)

To return to the example once again, an alternative to the cohort design would be to use a comparison group of students who were sixth graders in the same academic year 2017–2018 as were the students in the treatment group but, say, from a different school. As already noted, compared to that noncohort design, the advantage of the cohort design (as the dotted line suggests) is that the cohorts (because they are from the same, rather than different, schools) are likely to share many more features than the noncohort comparison group and hence have smaller selection differences. Another advantage of cohort designs is they do not require that the treatment be withheld from any participants at the time the intervention is implemented. A relative disadvantage of the cohort design is that cohorts are not as likely as noncohorts to experience the same history effects (see Section 7.11). For example, if different external events happened

during the 2016–2017 academic year than during the 2017–2018 academic year, this could impact the cohort design but not the noncohort design.

Both Bryk and Weisberg (1976) and Cook et al. (1975) used a modified cohort design based on age. In the case of Cook et al. (1975), the effect of *Sesame Street* was assessed by measuring a single group of preschoolers (of various ages) before *Sesame Street* was first aired and then again six months later after *Sesame Street* was aired. The children in the posttest sample were then matched for age with children from the pretest sample. For example, the posttest scores of children who were 5 years old at the time of the posttest were compared to the pretest scores of children who were 5 years old at the time of the pretest (but could not have seen *Sesame Street* at that point in time). That is, the counterfactual for estimating the treatment effect was derived from children who were the same age as the treatment group but were the same age at a different time of measurement.

7.11.3 Multiple Comparison Groups

I have already touched on the potential advantages of using two or more comparison groups for deriving tests of ignorability (see Section 7.9.4), but it is worthwhile to say a few words more. A multiple comparison group design would be diagrammed thusly:

$$\begin{array}{lccc} \text{NR:} & O_1 & X & O_2 \\ \hline \text{NR:} & O_1 & & O_2 \\ \hline \text{NR:} & O_1 & & O_2 \end{array}$$

To the extent that selection differences vary in each of the two comparisons of the treatment group with a comparison group and yet the estimates of treatment effects are the same, the credibility of the estimates of treatment effects are enhanced and the plausibility of the results being explained by hidden bias are reduced (Rosenbaum, 2017). For example, Roos, Roos, and Henteleff (1978) used two different comparison groups in their study of the effects of tonsillectomies on health outcomes. One comparison group was constructed in the usual way by selecting children who were like the children in the treatment group in terms of preexisting conditions but did not receive the treatment. The other comparison group consisted of untreated siblings. Both comparisons evidenced the same beneficial effects of tonsillectomies. To discredit these results, a critic would have to argue that biases due to the effects of selection differences were similar in both comparisons, which is not as plausible as assuming that the similar results are due to the shared treatment effect (also see Chapter 12).

Rosenbaum (2017) gave other examples of designs with multiple comparison groups, including a study of the effects of a change in health care policies on subsequent use of mental health services for federal employees. One comparison group consisted

of federal employees on whom data were collected the year before the policy was implemented. The other comparison group consisted of nonfederal employees on whom data were collected during the same year as the policy change. The contrasts of the treatment group to each of these two comparison groups yielded the same conclusions: the policy change reduced costs for subsequent mental health services. These two comparisons provide more credible evidence in favor of a treatment effect than does a single comparison; for further discussion, see Chapter 12.

Campbell (1969a) provided examples of the use of multiple comparison groups that ruled out threats to internal validity through "supplemental variation" wherein the researcher purposely degraded the quality of some of the comparisons. As Rosenbaum (2017, p. 151) explained:

> If the treatment was introduced at midnight on Tuesday, perhaps the best comparison is of treated Wednesday versus untreated Tuesday because trends over time that might confound the treatment effect are minimized; however, an additional comparison with Monday—an inferior control group further away in time—may help to show that no strong trend was present prior to the start of the treatment.

The same logic applies to a study where differences in motivation, say, might confound the comparison of nonequivalent groups (Rosenbaum, 2017). In this case, the single best comparison group might be one in which differences in motivation were minimized. But adding a second comparison group where motivation was substantially different (either stronger or weaker) could help rule out motivation as a threat by showing it was not the source of substantial differences in outcomes. More will be said about the strategies of adding data to rule out threats to validity in Chapter 11. Multiple comparison groups can also be used to bracket the size of a treatment effect within a range of plausible values (Section 13.9).

7.11.4 Multiple Outcome Measures

As already noted in Section 7.9.4 (also see Section 6.4), nonequivalent dependent variables can be added to a nonequivalent group design to perform a test of ignorability. A nonequivalent dependent variable can also be added to rule out alternative explanations, such as differential history effects. Remember, a nonequivalent dependent variable is a variable that is not expected to be influenced by the treatment but rather is expected to be susceptible to the same threats to validity as the intended dependent measure. A nonequivalent group design with a nonequivalent dependent variable would be diagrammed thusly:

$$\begin{array}{llllll} \text{NR:} & O_{1A} & O_{1B} & X & O_{2A} & O_{2B} \\ \text{NR:} & O_{1A} & O_{1B} & & O_{2A} & O_{2B} \end{array}$$

where O_A and O_B are the nonequivalent measures. Suppose measure O_A was expected to show the effects of the treatment, while measure O_B was not. For example, the treatment might be a medical intervention in a hospital expected to influence one illness but not another where measures O_A and O_B assessed outcomes relevant to one but not the other illness. Further suppose that the treatment group, but not the comparison group, was susceptible to a differential history effect and that the differential history effect was expected, if it had an effect at all, to equally affect both measures O_A and O_B. For example, perhaps some of the medical staff at the experimental hospital had been upgraded between the time of the pretest and posttest. To rule out a bias due to differential history, the researcher would analyze the data from measure O_B in the same way the data from measure O_A would be analyzed to estimate the treatment effect. That is, the researcher would look for a pseudo treatment effect with the data from the nonequivalent dependent variable. A null result with measure O_B would indicate not only that no treatment effect was present (as presumed) but also that there was no effect of differential history. Hence, the analysis of the data using measure O_A would be presumed to be free from the effects of differential history as well.

7.11.5 Multiple Pretest Measures over Time

Instead of having multiple (nonequivalent) outcome measures, another option would be to have multiple pretest measures spaced out over time. Such a double pretest design would be diagrammed thusly:

NR: O_1 O_2 X O_3

NR: O_1 O_2 O_3

The additional pretest can be used in three ways. First, the double pretests could provide an additional test of ignorability (see Section 7.9.4). The data from the two pretests would be analyzed as if they had come from a pretest and a posttest (e.g., using ANCOVA or matching/blocking with or without propensity scores). The result should be a finding of no treatment effect because no treatment was introduced between the time of the first and second pretests. Such a null finding would suggest (though not prove) that the analysis model used to take account of selection differences in analyzing the second wave of pretest data as the outcome variable would similarly work to take account of selection differences in analyzing the true posttest data as the outcome variable. That is, if an analysis worked with waves of data 1 and 2, the same analysis would be presumed to work for waves 2 and 3. Boruch (1997; Boruch & Gomez, 1977) called such a test of ignorability a "dry-run" analysis.

Second, when analyzing the posttest data using ANCOVA or matching/blocking, researchers could use data from both the pretests to take account of selection

differences, thereby helping to model differences between the groups in growth patterns over time.

Third, the double pretests could be used to predict growth over time in an extension of a change score or DID analysis (see Section 7.3). That is, the pattern of growth from the first to the second pretest could be extrapolated to predict growth from the second pretest to the posttest. For example, if the mean difference between the treatment groups increased from the first to the second pretest, a similar increase would be predicted to occur by the time of the posttest, in the absence of a treatment effect. Differences between the predicted change and the obtained change in the mean posttest scores would be attributed to the effect of the treatment. Wortman et al. (1978) provide an example of such an analysis. Across two waves of pretest observations, the treatment group was growing progressively further ahead of the comparison group. After the treatment was introduced, the performance in the treatment group lagged the performance in the comparison group. Especially given the two waves of pretest observations, such a pattern of change in group means over time could not be plausibly explained as anything other than a negative treatment effect.

7.11.6 Multiple Treatments over Time

Another way to supplement a nonequivalent group design is to add treatment variations over time along with added times of measurement. Consider three such supplements.

The first supplemented nonequivalent group design adds a treatment variation by either removing or reversing the treatment after it has once been introduced. The design is diagrammed thusly:

NR: O_1 X O_2 $\not{X}$ O_3

NR: O_1 O_2 O_3

where $\not{X}$ denotes either the removal or reversal of the treatment. In other words, the treatment is introduced and then either removed or, if possible and appropriate, reversed.

The second supplemented nonequivalent group design entails adding a treatment to the comparison condition but after the treatment is introduced in the original treatment group. The design is diagrammed thusly:

NR: O_1 X O_2 O_3

NR: O_1 O_2 X O_3

The treatment is introduced in one treatment condition between the first and second measurements and is introduced to the other treatment condition between the second

and third measurements. Such a design is called a nonequivalent group design with **switching replications** (Shadish et al., 2002).

The third variation is a crossover design where different treatments are introduced at different time points. It is diagrammed thusly:

$$\begin{array}{llllll} \text{NR:} & O_1 & X_A & O_2 & X_B & O_3 \\ \hline \text{NR:} & O_1 & X_B & O_2 & X_A & O_3 \end{array}$$

where X_A and X_B are alternative treatments presumed to have different patterns of effects.

Each of these designs allows the estimate of the treatment effect to be replicated in one form or another. A particularly simple way to analyze data from such designs is to consider changes in the mean outcomes over time as an extension of the change score or DID analyses (see Section 7.3). From this perspective, the purpose of the design is to introduce a pattern of results that is difficult to explain as being due to selection differences. Consider the switching replication design. The treatment effect should cause varying differences between the groups over time. That is, if the treatment has an effect, the treatment should cause the gap between the two treatment groups to shift in one direction between the times of the first and second measurements and in the opposite direction between the times of the second and third observations. For example, if the first group starts out ahead of the second group and if the treatment effect is positive, the gap between the treatment groups should increase by Time 2 and then diminish by Time 3. In contrast, the nature of effects due to selection differences is presumed to remain the same over time. If the results turn out to be in opposite directions as predicted by the treatment effects, the results are difficult to explain as being due to selection differences.

7.12 EMPIRICAL EVALUATIONS AND BEST PRACTICES

Can the nonequivalent group design produce unbiased (or at least adequate) estimates of treatment effects in practice? A substantial literature exists that attempts to answer that question (Shadish, 2000). In this literature, the results of randomized experiments are presumed to provide unbiased estimates of treatment effects. These presumably unbiased estimates are compared to the treatment effect estimates provided by nonequivalent group designs. The results of the nonequivalent group designs are judged adequate to the extent that they agree sufficiently well with the presumed unbiased estimates from the randomized experiments.

Such comparative studies fall into two categories: between-study and within-study designs. The classic between-study design is a **meta-analysis** that collects the results from independent studies on the same substantive topic and compares the aggregated

results of randomized experiments to the aggregated results of nonequivalent group designs (e.g., Lipsey & Wilson, 1993). In contrast, in the classic within-study design, a randomized experiment is compared to a yoked nonequivalent group design where both designs share the same treatment group but use different comparison groups. For example, a randomized experiment is implemented with randomly equivalent treatment and comparison conditions labeled group A and group B, respectively. Then a nonequivalent comparison condition (group C) is added. The treatment effect estimate from the randomized experiment is derived from a comparison of groups A and B. The treatment effect estimate from the nonequivalent group design is derived from a comparison of groups A and C. Cook, Shadish, and Wong (2008) provide guidelines for conducting state-of-the-art within-study comparisons.

The results from both types of designs have been variable. Some studies find that randomized experiments and nonequivalent group designs produce similar results. Other studies find that randomized experiments and nonequivalent group designs produce dissimilar results. Less equivocal are conclusions being drawn from such studies about the circumstances that produce the best estimates from nonequivalent group designs. In other words, the variable outcomes from within- and between-study designs are leading to empirically justifiable conclusions about best practices for the implementation of nonequivalent group designs (Coalition for Evidence-Based Policy, 2014; Cook et al., 2008; Heinsman & Shadish, 1996; Shadish, 2000). What follows are four conditions that appear to produce the most credible results when using nonequivalent group designs (Cook et al., 2008).

7.12.1 Similar Treatment and Comparison Groups

The treatment and comparison groups should be similar right from the start, especially on important measures such as pretests that are operationally identical to the posttests and on other characteristics that greatly influence either outcomes or selection into treatment conditions. It is better to avoid pretreatment differences as much as possible than to try to remove them with statistical machinations because statistical machinations cannot be counted on to do so. For example, rather than comparing eligible participants to ineligible participants, in some cases it might be better to compare those who chose to enroll in a program to those who were eligible but declined or were unable to enroll in the program. Another way to make treatment and comparison groups similar at the start is to draw the treatment groups from nearby rather than distant geographical locales. For example, in a study of the effects of an educational innovation, groups of participants should be drawn from the same rather than different school districts.

Comparison groups that are like treatment groups in both location and pretreatment characteristics have been called **focal local comparison groups** (Steiner et al., 2010). Cohort designs (see Section 7.11.2) are one way to produce these groups. In addition, in making treatment groups similar, it is also important to avoid noncompliance and missing data as much as possible.

7.12.2 Adjusting for the Selection Differences That Remain

Researchers should adjust rigorously for remaining selection differences after using focal local comparison groups. Whether using matching, regression adjustments, or some other techniques, researchers need to implement these strategies with care and forethought. When using matching, researchers need to create close matches between participants in the treatment groups. When using regression techniques, researchers need to adjust for nonlinear relationships and interactions. All the different adjustment methods rest on stringent assumptions. No methods will automatically take account of selection differences regardless of the research circumstances. Researchers need to be aware that violating the underlying assumptions can lead to severe biases in treatment effect estimates and, therefore, need to be attuned to satisfying the assumptions of the chosen analytic methods. Researchers also need to recognize that all the adjustment procedures are implemented best with large sample sizes. For example, instrumental variable methods are biased and will only converge on unbiased results as the sample size increases. Matching methods (such as with propensity scores) can require large pools of participants from which to find adequate matches.

Rather than performing only a single analysis to adjust for the effects of selection differences, multiple methods should be used. To the extent that the multiple methods rely on different assumptions about how to take account of selection differences and to the extent that the results from the different analyses agree, the credibility of the results is increased.

7.12.3 A Rich and Reliable Set of Covariates

As already noted in Section 7.4.1, no method of adjusting for selection differences will work well without adequate covariates on which to perform the adjusting. Indeed, it has been argued that the method of adjustment is not nearly as important as the pool of available covariates (Angrist & Pischke, 2009; Bloom, 2005a; Cook & Steiner, 2010; Shadish, 2013). Research suggests that it does little good to adjust for standard demographics such as sex, age, marital status, and race (Shadish et al., 2008). That is, standard demographic measures appear to produce little reduction in bias due to selection differences. Much more important is a rich set of covariates that can be used to model either the outcomes or the selection process in the particular research setting (see Section 7.4.1).

For modeling outcomes, the greater the correlation between covariates and outcomes, the better is the model in general. Often the single best covariate for modeling outcomes is a covariate that is operationally identical to the posttest—or a covariate that is a close proxy for the posttest (Cook & Steiner, 2010, Cook et al., 2009, Glazerman, Levy, & Myers, 2003; Wong et al., 2012). Note, too, that as the time lag between pretest and posttest decreases, the correlation between operationally identical pretests and posttests tends to increase, which generally means better bias reduction.

For modeling the process of selection, there is no substitute for knowledge of the basis on which participants are selected into groups. If participants are selected based on certain characteristics, make sure to measure those characteristics and include them in the statistical model. For example, if the more motivated participants self-select into the treatment group, measure motivation for enrolling in the treatment. Note that for modeling selection, it appears best not to allow individual participants to self-select into treatment conditions but to allow selection only by other methods, such as the selection of intact groups or selection by administrators (Heinsman & Shadish, 1996).

Also remember that ANCOVA and matching/blocking are biased by the presence of measurement error in the covariates, so it is important to have covariates that are measured reliably. In this regard, note that measures of intact groups are often more reliable than measures of individual participants. For example, school-level measurements aggregated from individual student-level measurements tend to be far more reliable than measurements of individual students. In addition, composite measures and measures that are aggregates of multiple pretests over time tend to be more reliable than individual measures. In addition, when selecting treatment groups to be similar (see Section 7.12.1), select them based on more reliable and stable characteristics, such as school-level measurements rather than individual student-level measurements.

7.12.4 Design Supplements

Section 7.11 described supplements that can be made to the basic pretest–posttest nonequivalent group design. Chapters 11 and 12 provide more details about the logic and practice of adding design elaborations. Such supplements can both increase power and precision and help rule out threats to internal validity (including biases such as those due to selection differences and differential history effects) and thereby increase the credibility of results.

7.13 STRENGTHS AND WEAKNESSES

Nonequivalent group designs are some of the most common quasi-experiments but often produce less credible results than other, more advanced quasi-experimental designs (see Chapters 8 and 9). Nonequivalent group designs are common because they are easy to implement, and they are one of the less credible designs because of selection differences. Selection differences are likely to be present in nonequivalent group designs and can bias the estimates of treatment effects. Unfortunately, there is no guaranteed method for taking account of selection differences in a nonequivalent group design once they are present. All methods for taking account of selection differences rely on assumptions that are ultimately unverifiable. Researchers must rely on the best information

that is available, while recognizing that information is inevitably inadequate. Nonetheless, researchers can create, with diligence and forethought, some nonequivalent group designs that are better than other nonequivalent group designs.

Some methods of adjusting for selection differences are more widely used and appear to produce more credible results than others. ANCOVA and matching/blocking with or without propensity scores appear to be most common. Instrumental variable (IV) approaches are also common, especially in sociology and economics (Angrist & Pischke, 2009). But my sense is that conditions conducive to IV methods are far less common than conditions conducive to ANCOVA and matching/blocking methods. (Economists and sociologists often appear to prefer IV methods.) Change-score analyses (especially those under the guise of DID analyses) also appear to be common adjustment methods, though the underlying assumptions of these methods are strict and limit their plausible application as much as do the assumptions of IV approaches. Selection modeling approaches appear to be the least frequently used, having fallen out of favor since their heyday a decade or so ago. Sensitivity analysis is not used as often as it should be, and the same can be said for design supplements.

It is difficult to know which statistical analysis procedures are superior to others in any given set of research circumstances. The best approach is to use multiple methods. To the extent their assumptions appear varied but justified, and to the extent that the results of the multiple analyses (including sensitivity analyses) agree, researchers can be more confident in the estimates of treatment effects (Ho et al., 2007). Conversely, to the extent that the results of multiple plausible analyses disagree, researchers will need to be correspondingly circumspect about the size of treatment effects. It is best to follow the advice given in Section 7.12. Choose nonequivalent groups carefully so that they are as similar at the start as possible. Be diligent in using statistical methods to take account of the selection differences that remain; collect a thorough and reliable set of covariates for modeling outcomes and/or selection processes; and use design supplements.

Besides ease of use, the relative advantages of nonequivalent group designs lie mostly in their external validity. That nonequivalent group designs are easy to implement means they can be implemented under conditions in which other designs cannot be implemented. For example, many administrators would be more receptive to nonequivalent group designs than to randomized experiments. Hence, the results of nonequivalent group designs can often be generalized to a wider range of circumstances than can the results of randomized experiments.

If, however, your focus is on internal validity more than external validity (as I believe it usually should be), the nonequivalent group design tends to be less useful than randomized and other more credible quasi-experimental designs, such as the regression discontinuity design and the interrupted time-series design (see Chapters 8 and 9). Of course, a randomized experiment can be degraded by noncompliance and differential attrition, so that a randomized experiment becomes a broken randomized experiment. But the biases introduced by noncompliance and differential attrition in

randomized experiments are often less severe than the biases due to selection differences in nonequivalent group designs.

7.14 CONCLUSIONS

Nonequivalent group designs are relatively easy to implement—much easier in fact than between-groups randomized experiments. At the same time data from nonequivalent group designs are tricky to analyze. As Schafer and Kang (2008, p. 280) remark, "Even under the assumptions of unconfoundedness, causal inference is not trivial; many solutions have been proposed and there is no consensus among statisticians about which methods are best." As a result, nonequivalent group designs tend to produce far less credible results than randomized experiments. According to the What Works Clearinghouse (WWC), even the very best nonequivalent group design is not eligible to receive their highest rating of "Meets WWC Group Design Standards without Reservations" (U.S. Department of Education, 2017). The best a nonequivalent group design can do is "Meet WWC Group Design Standards with Reservations." Huitema (1980, p. 352) provides an accurate, if gloomy, summary when he states that "nonequivalent-group designs are very weak, easily misinterpreted, and difficult to analyze." Other designs are often to be preferred. With adequate care and effort in implementation, however, the nonequivalent group design can be cautiously used to good effect, especially when other designs are not feasible.

7.15 SUGGESTED READING

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27,* 724–750.

—Describes conditions under which nonequivalent group designs are most likely to produce unbiased estimates of treatment effects.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York: Cambridge University Press.

—Provides further technical details on designing and analyzing data from nonequivalent group designs.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Cambridge, MA: Harvard University Press.

—Provide further technical details on designing and analyzing data from nonequivalent group designs.

May, H. (2012). Nonequivalent comparison group designs. In H. Cooper (Ed.), *The APA handbook of research methods in psychology* (pp. 489–509). Washington, DC: American Psychological Association.
—Provides an easy-to-follow summary of the nonequivalent group design.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.
—Gives an overview of matching methods.

Introductions to propensity score analysis with many details that go beyond what is presented in this chapter can be found in the following articles, chapters, and books:

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*, 399–424.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22,* 31–72.

Li, M. (2012). Using propensity score method to estimate causal effects: A review and practical guide. *Organizational Research Methods, 16,* 188–226.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29,* 530–558.

Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Cambridge, MA: Harvard University Press.

Steiner, P. M., & Cook, D. (2015). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 1, pp. 237–259). New York: Oxford University Press.

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and applications in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology, 82,* 906–919.

8

Regression Discontinuity Designs

The implication is that [the RD design] is a robustly effective tool that mere mortals can use if they have considerable, but not necessarily perfect, sensitivity to its assumptions.

—Cook, Shadish, and Wong (2008, p. 732)

Researchers who need to use [a RD design] can . . . have reasonable confidence that they are getting an accurate estimate of the effects of treatments.

—Shadish, Galindo, Wong, Steiner, and Cook (2011, p. 190)

Some American government agencies that commissioned evaluations began to specify that [RD designs] should be used in preference to other methods if [a randomized] experiment was not possible. . . .

—Cook and Wong (2008b, p. 132)

Overview

The regression discontinuity design requires that participants be assigned to treatment conditions based on a cutoff score on a quantitative variable (called the quantitative assignment variable [QAV]). The QAV can be a measure of need or merit, for example, so that the neediest or most meritorious participants receive the treatment, while the less needy and less meritorious are assigned to a comparison condition. Such an assignment to treatment conditions can be appealing under certain circumstances, such as when stakeholders want all the neediest or most meritorious to receive the treatment.

In the analysis of data from the regression discontinuity design, the outcome scores are regressed onto the QAV in each treatment group. The effect of the treatment is estimated as a discontinuity (either in level or slope of the regression surfaces) across the treatment groups at the cutoff score. A primary threat to the internal validity of the regression discontinuity design is that the regression surfaces are not modeled correctly because of curvilinearity. A variety of statistical techniques have been advanced for correctly modeling the regression surfaces. Other threats to internal validity such as those due to noncompliance, attrition, or manipulation of the QAV can also arise.

8.1 INTRODUCTION

The regression discontinuity (RD) design was first invented by Thistlewaite and Campbell (1960). I say "first" invented because the design was later reinvented several times, apparently independently (Cook, 2008b; Shadish et al., 2002). Since those early days, the design has been a mainstay in texts on applied research methods in social sciences, such as psychology and education (Reichardt & Henry, 2012). Nonetheless, the design has had uneven use over the years. The design was used in the 1970s to assess the effects of compensatory educational programs funded by Title I of the 1965 Elementary and Secondary Education Act (Trochim, 1984). The design did not become popular again until the 1990s when it was rediscovered and put to good use, especially by economists (Cook, 2008b). For example, Lee and Lemieux (2008) list over 70 studies on economic-related topics that have used the RD design in relatively recent years. The RD design has also received increasing attention from funding agencies such as the Institute for Educational Sciences of the U.S. Department of Education. Although many consider the randomized experiment to be the only gold standard for estimating the effects of treatments, the RD design is coming to be considered a respectable substitute (and even a gold standard by some), especially when a randomized experiment is not acceptable to administrators, staff, or participants (Cook & Wong, 2008a; Shadish et al., 2011; Somers, Zhu, Jacob, & Bloom, 2013; Sparks, 2010).

In the basic RD design, participants are assigned to treatment conditions based solely on a variable, called the **quantitative assignment variable (QAV),** upon which each participant is assessed. (In some literatures, the QAV is called the forcing or running variable.) Specifically, a cutoff score on the QAV is specified, and participants are assigned to treatment conditions based on a comparison between their QAV score and the specified cutoff score. Those participants with QAV scores above the cutoff score are assigned to one treatment condition, and those with QAV scores below the cutoff score are assigned to the other treatment condition. (Those with scores equal to the cutoff score are assigned to one or the other of the treatment groups, depending on a specified decision rule.) After participants are assigned to treatment conditions, the treatments are implemented, and, after the treatments have been given time to have their effects, the participants are assessed on one or more outcome measures.

The effect of the treatment is assessed in the following way. The scores on a given outcome variable are regressed onto the QAV scores in each of the two treatment groups, and these two regression lines are compared. The effect of the treatment is estimated as a discontinuity, between the two regression lines (which gives the design its name). This can best be understood with figures. (If there are two or more outcome variables, each can be treated in the same way.)

Figures 8.1, 8.2, 8.3, and 8.4 are scatterplots of data from four hypothetical RD designs. In each figure, the scores on the outcome variable are plotted along the vertical axis, and the QAV scores are plotted along the horizontal axis. The cutoff score is 40 in each case and is marked with a vertical dashed line in each figure. The participants

with QAV scores less than the value of 40 were assigned to the treatment condition, while those with QAV scores equal to or greater than the value of 40 were assigned to the comparison condition. The scores for the participants in the treatment condition are marked with small squares in the figures, and the scores for the participants in the comparison condition are marked with small circles. The sloping solid lines in each figure are the regression lines of the outcome scores regressed on the QAV scores, estimated separately in each of the treatment conditions.

In Figure 8.1, the two regression lines are the same—the two lines have the same intercept and slope. In particular, there is no discontinuity between the two lines at the cutoff score. As a result, the treatment would be estimated to have no effect.

In contrast, the two regression lines are not the same in Figure 8.2. The regression line in the treatment condition is shifted vertically compared to the regression line in the comparison condition (the solid lines in the figure). As a result, there is a displacement, or discontinuity, between the two regression lines at the cutoff score. In the data in Figure 8.2, the treatment would be estimated to have a positive effect because the regression line in the treatment group is displaced upward compared to the regression line in the comparison condition. If the regression line in the treatment group had been displaced downward compared to the regression line in the comparison condition, the treatment would be estimated to have a negative effect. Also note that the treatment appears to have a positive effect at all levels of the QAV. That is, if you projected the two regression lines across the entire range of QAV scores (shown by the dashed lines in the figure), the regression line for the treatment condition would be elevated above the regression line for the comparison condition everywhere along the QAV.

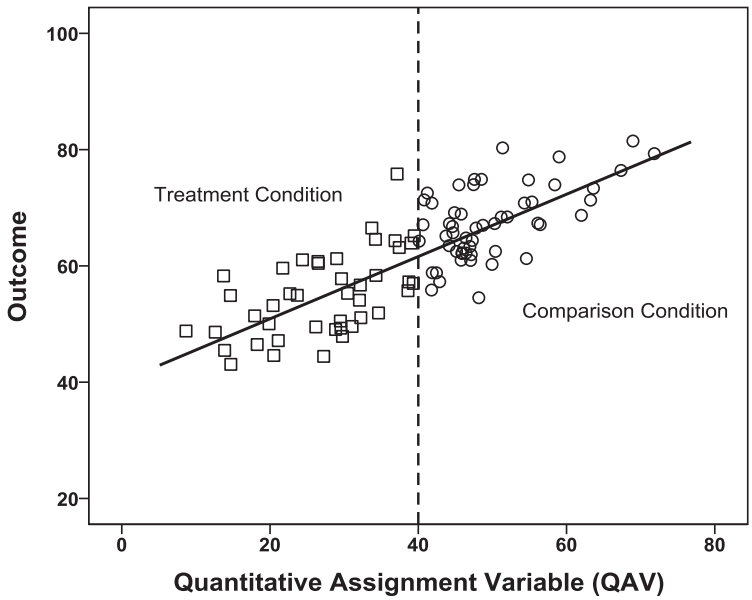


FIGURE 8.1. Hypothetical results from a regression discontinuity design with no treatment effect.

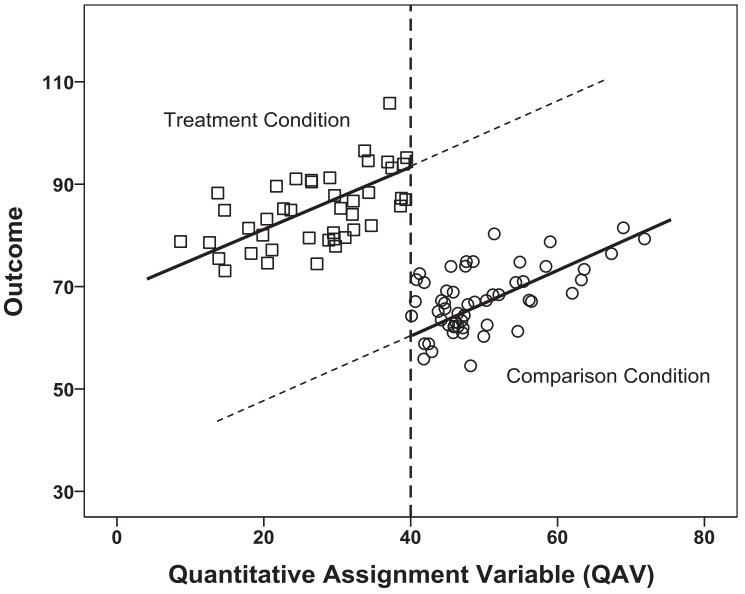


FIGURE 8.2. Hypothetical results from a regression discontinuity design where a treatment effect produces a change in level at the cutoff score.

In Figure 8.3, the two regression lines are not displaced at the cutoff score on the QAV relative to one another, but they have different slopes. If the regression lines are projected onto each side of the cutoff score (as shown by dashed lines in the figure), the treatment would be estimated to have an increasingly positive effect as scores on the QAV decrease from the cutoff score. That is, on the left side of the cutoff score, the regression line from the treatment group is farther and farther above a projected regression line from the comparison condition, as scores on the QAV decrease. On the other hand, the treatment would be estimated to have a negative effect on participants with QAV scores above the cutoff score because a projected regression line from the treatment condition (the dashed line for the treatment group) lies below the regression line in the comparison condition. Because of the nonparallel regression lines, there is said to be an interaction effect between the treatment and the QAV. The treatment is said to cause a discontinuity in the slope of the regressions.

In Figure 8.4, the two regression lines are not only displaced vertically compared to one another, but also they are tilted compared to each other. In other words, not only is there a displacement or discontinuity in level between the two regression lines at the cutoff score, but there is also a discontinuity in regression slopes in the two treatment conditions. In this case, the treatment would be estimated to have a positive effect at the cutoff score on the QAV and to have a different size effect at other QAV scores. That is, the treatment has an effect at the cutoff score, and the effect of the treatment also interacts with the QAV scores. In Figure 8.4, the treatment is estimated to have a larger effect for participants with high QAV scores than for participants with low QAV scores.

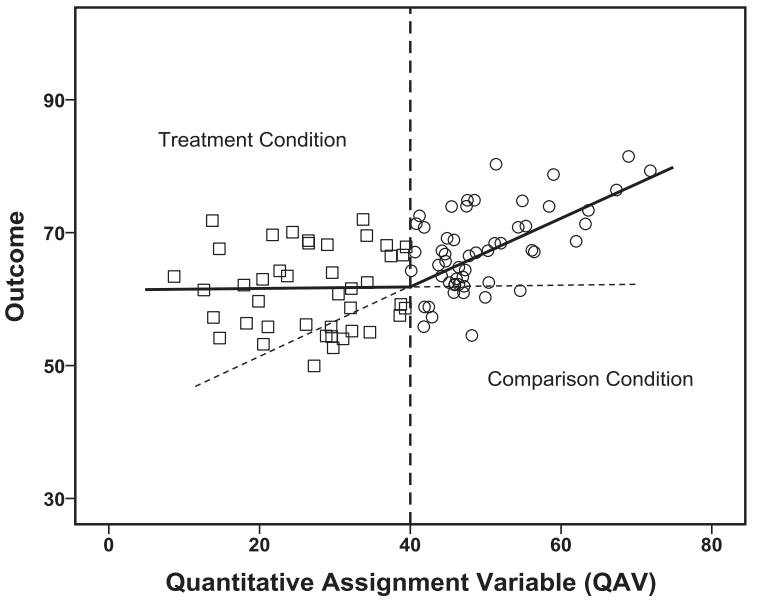


FIGURE 8.3. Hypothetical results from a regression discontinuity design where a treatment effect produces a change in slope but not a change in level at the cutoff score.

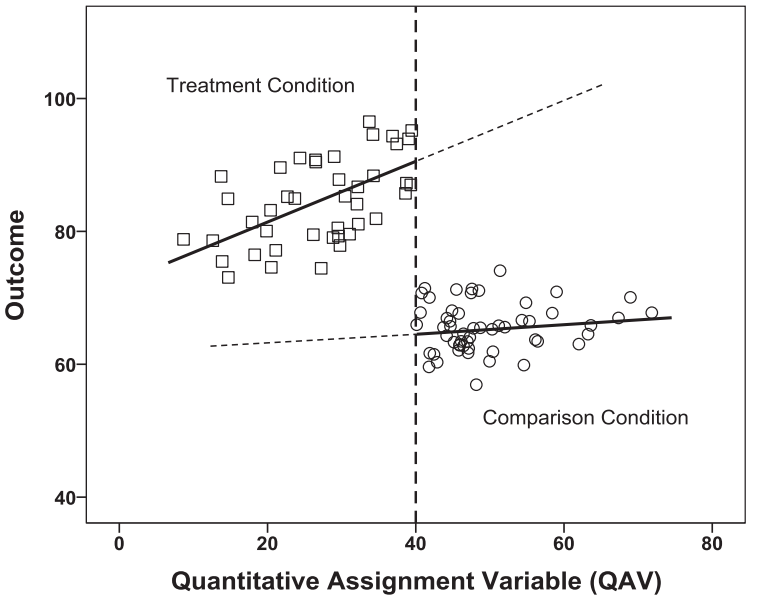


FIGURE 8.4. Hypothetical results from a regression discontinuity design where a treatment effect produces both a change in level at the cutoff score and a change in slope.

That is, if you projected the two regression lines across the entire range of QAV scores (the dashed lines), the treatment would appear to have raised the outcome scores more for those with high QAV scores than for those with low QAV scores.

A treatment that causes one regression line to be displaced above or below the other regression line at the cutoff score is said to have caused a change in level at the cutoff score. A treatment that causes a change in the tilt of one regression line compared to the other is said to have caused a change in slope. These two effects are orthogonal. As Figures 8.1–8.4 reveal, you can have neither of the two effects, one or the other of the effects, or both effects.

As described below, a researcher seeks to implement an RD design under circumstances where, if the treatment has no effect, the two regression lines would fall on top of one another, as in Figure 8.1. Then discrepancies between the regression lines, such as in Figures 8.2, 8.3, and 8.4, would be due to the effects of the treatment.

It can be instructive to compare data from an RD design to data from a randomized experiment. Figure 8.5 presents data that might have been obtained if the RD design depicted in Figure 8.4 had been conducted as a randomized experiment instead. In Figure 8.5, the data for the treatment group extend on both sides of the cutoff score on the QAV from the RD design, as do the data for the comparison group. It is easy to see in Figure 8.5 that the treatment has both an effect of a change in level at the cutoff score and a change in slope (i.e., an interaction effect). The effect of a change in level at the cutoff score is such that the average of the outcome scores in the treatment group lies above the average of the outcome scores in the comparison group at the cutoff score

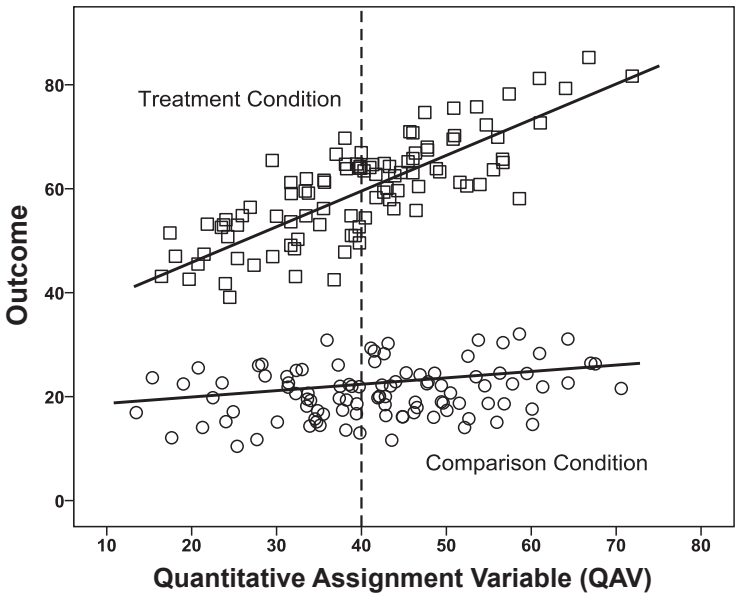


FIGURE 8.5. Hypothetical results from a randomized experiment used to demonstrate the difference between a randomized experiment and a regression discontinuity design.

on the QAV. The effect of a change in slope is such that the vertical displacement of the regression line in the treatment group is greater for participants with high QAV scores than for participants with low QAV scores. Comparing Figure 8.4 to Figure 8.5, it can be seen that an RD design is the same as a randomized experiment. The difference is that part of the data in each treatment condition is missing in the RD design. Nonetheless, the two designs estimate the same treatment effects.

In design notation, the RD design is diagrammed thusly :

$$\begin{array}{llll} \text{C:} & O_1 & X & O_2 \\ \hline \text{C:} & O_1 & & O_2 \end{array}$$

As before, O denotes an observation, and X denotes the implementation of the treatment. The first observation (O_1) is the QAV, and the second observation (O_2) is the outcome variable. The "C:" in the diagram denotes that assignment has been based on a cutoff score on the QAV.

8.2 THE QUANTITATIVE ASSIGNMENT VARIABLE

The quantitative assignment variable can be a measure of need or merit, so either the neediest or the most meritorious participants are selected to receive the treatment condition, rather than the comparison condition. Alternatively, the quantitative assignment variable can be based on other qualities such as age or time of application for treatment. We will now consider each type of QAV in turn.

8.2.1 Assignment Based on Need or Risk

If the treatment is meant to alleviate a problem or ameliorate a deficit, the QAV can be a measure of the severity of that problem or deficit. The treatment would then be given to the participants with the greatest need or deficit (or those at greatest risk of suffering loss in the absence of the treatment) and the participants with less need, deficit, or risk would serve in the comparison condition.

For example, Jacob and Lefgren (2004) used an RD design to assess the effects of a program of remedial instruction during the summer months on the reading and math achievement of third- and sixth-grade students. The QAV was a measure of cognitive skills assessed at the end of the academic year before the summer vacation. A low score on the test of cognitive skills led to placement in the summer school program.

Buddelmeyer and Skoufas (2004) provide another example of an RD design with an ameliorative treatment intervention. They conducted an RD design to assess the effects of cash payments where the QAV was family wealth, with low scores making the family eligible to receive the treatment. The outcomes were school attendance and health care.

Or consider Aiken, West, Schwalm, Carroll, and Hsiung (1998). They assigned first-year incoming college students to remedial English classes based on scores on the verbal subtest of the SAT. Students with low scores were enrolled in the remedial program, and the outcome measure was an assessment of writing ability.

8.2.2 Assignment Based on Merit

Alternatively, the treatment could be a program or an intervention that rewards meritorious performance, in which case the QAV could be a measure of merit or ability, with the treatment given to those exhibiting the highest degree of merit or ability. For example, Thistlethwaite and Campbell (1960) assessed the effects of winning a National Merit Scholarship Certificate of Merit on both the chance of earning a college scholarship and career aspirations. The QAV was performance on the National Merit Scholarship qualifying test, and the cutoff score was the score required to earn the Certificate of Merit. Those with sufficiently high scores on the qualifying test received the Certificate of Merit.

Van der Klaauw (2002) assessed the effects of an offer of financial aid on enrollment in college. The QAV was academic ability, as measured by a composite variable that included SAT scores and high school grades. The cutoff score was based on the specifications colleges used to award financial aid. Those high school students with high scores on the QAV received offers of financial aid.

Seaver and Quarton (1976) assessed the effects of being placed on the Dean's List during the fall quarter in college on grades the next quarter. The QAV was fall quarter grade point average (GPA) where the cutoff score was a GPA of 3.5. Seaver and Quarton concluded that being placed on the Dean's List improved grades in the subsequent quarter, though this result has subsequently been called into question (Shadish et al., 2002).

Berk and Rauma (1983) assessed the effects of a program in California where unemployment compensation was given to prisoners upon being released from prison. The QAV was the number of hours the prisoners had worked during the last year in prison. The cutoff score was 652 hours, with prisoners who worked more hours given unemployment compensation. The outcome measure was recidivism, which was estimated to be reduced by 13%. The program became part of California law.

8.2.3 Other Types of Assignment

Although either need or merit is the basis for the most common QAV measures, the QAV need not be based on either. Other criteria for assigning participants to treatment conditions could be and have been used. For example, participants could be assigned to treatments based on the time at which they apply or arrive for treatment, so the first to apply or arrive are those who receive the treatment (Trochim, 1984). Consider other options as well.

Lee (2008) assessed the effects of incumbency on the election of Democrats (versus Republicans) to the U.S. House of Representatives. The QAV was the proportion of the vote given to the Democratic candidate in the prior election. The participants were those who ran for election two terms in a row. Those with greater than 50% of the vote during the first term were obviously the incumbent. The effect of incumbency was estimated to have a substantial effect on winning a subsequent election.

Cahan and Davis (1987) estimated the effects of the first year of schooling on academic achievement in verbal and math skills. The QAV was age, and the cutoff score was the minimum age required for enrollment in first grade (which was 6 years old by December 31). Outcomes were measured at the end of that first academic year, with the performance of the children who were old enough to enroll in first grade that year compared to the performance of children who had not been old enough to enroll in first grade that year. Joyce, Kaestner, and Colman (2006) also used a QAV of age to assess the effects of a Texas law implemented in 2000 requiring that parents be notified of a request for an abortion by a minor child but not for those 18 or older.

Or consider DiNardo and Lee (2004). They assessed the effects of unionization on the closing of businesses. The QAV was the percentage of the vote in favor of unionization, with the cutoff score being a majority vote in favor of unionization. For one last example, Black (1999) used geographical location of school boundaries as a QAV.

8.2.4 Qualities of the QAV

The QAV can be either a set of ordered categories or a continuum of scores, but according to most sources, they must be at least at the level of an interval measurement. However, the What Works Clearinghouse (WWC) allows the QAV to be at the level of an ordinal measurement as long as there are at least four unique values on the QAV both above and below the cutoff score (U.S. Department of Education, 2017). In any case, the QAV cannot be a nominal measurement such as sex or race. The point is that the QAV must take on numerous interval-level/ordinal-level values so that regression lines can be reasonably fit to the data. The power and precision of the RD design are generally maximized when there are an equal number of participants on each side of the cutoff score (Cappelleri, 1991).

The QAV need not be related to the outcome measure—the logic of the design still applies even if the correlation between the QAV scores and the outcome scores is zero. Although there need not be a correlation between the QAV and the outcome variable, the estimate of the treatment effect will be more precise (and a test of the statistical significance of the treatment effect estimate will be more powerful) the higher is the correlation (in absolute value) between the QAV and the outcome measure. Because operationally identical measures are often most highly correlated, an effective QAV measure is often operationally identical to the outcome measure. For example, using the same test of cognitive abilities for both the QAV and the outcome measures would make the two measures operationally identical.

The primary requirement for the QAV is that the regression surfaces between outcome and QAV have the same (linear or curvilinear) shape (with no discontinuity) in both treatment conditions in the absence of a treatment effect. Preferably, those regression surfaces have a simple shape such as a straight line as in Figures 8.1–8.4. (In Section 8.3, I say more about what happens when the regression surface is curvilinear.)

The QAV can be derived from a single measure or from an aggregate of separate measures. For example, in their RD design, Henry, Fortner, and Thompson (2010) assigned schools to treatment conditions using a composite derived from four different measures—two based on teacher characteristics (average teacher longevity and experience) and two based on student characteristics (percent living in poverty and percent proficient in academic skills). The researchers combined the four measures into a single quantitative index, with a cutoff score used to assign schools to treatments. Some or all of the components that go into a QAV can even be subjective assessments as long as the subjective assessments are converted to a quantitative scale for use in the QAV. It is also possible to use multiple QAVs, which are not aggregated into a single measure (Cappelleri & Trochim, 2015; Papay, Willett, & Murnane, 2011; Wong, Steiner, & Cook, 2013).

Given that measurement error in a covariate can bias the estimate of a treatment effect in a nonequivalent group design (see Section 7.4.2), it is worth emphasizing that measurement error in the QAV will not introduce bias in the RD design. As long as the same QAV, however fallibly measured, is used both for assignment to treatment conditions and for estimate of the treatment effect in the statistical analysis, measurement error will not introduce any bias. For example, if cognitive ability is used as the QAV, cognitive ability need not be measured free of error. All that is required is that the same measure of cognitive ability be used to assign participants to treatment conditions, as is used to estimate the treatment effect. When the same measure is used in this fashion, the QAV is a perfectly reliable covariate in the analysis of data from an RD design because the QAV is a perfectly reliable measure of how participants have been assigned to treatment conditions (Cappelleri, Trochim, Stanley, & Reichardt, 1991; Trochim, Cappelleri, & Reichardt, 1991).

Sometimes participants trickle into the study, as when clients enter treatment over time rather than all at once (Braucht & Reichardt, 1993). In such cases, it can be difficult to establish a cutoff ahead of time to ensure that the treatment services become fully subscribed, as is often most desirable. In this case, it is possible to vary the cutoff score as the study proceeds but then, for the data analysis, to rescale the QAVs so that the cutoff scores are equated for different participants (e.g., by converting the QAV scores into the degrees of deviation from the different cutoff scores). The same strategy can be used if a researcher conducts a series of RD designs at multiple sites (Hallberg et al., 2013). The data could be combined into a single study where the cutoff scores have all been equated. That is, the data could be entered into the single analysis model where the QAVs from the different studies have been centered at their respective cutoff scores. Or the data from the multiple sites could be analyzed as individual studies and the results combined, using meta-analytic methods.

8.3 STATISTICAL ANALYSIS

Estimating a treatment effect in the RD design requires that the functional form of the relationship between the outcome variable and the QAV (i.e., the regression surface between outcome variable and QAV) is modeled correctly. Fitting the wrong regression surface can bias the estimates of the treatment effects. Consider Figure 8.6, which demonstrates, for example, how fitting straight lines in the presence of a curvilinear relationship can bias the estimates of the treatment effects. As in the previous figures, the cutoff score in Figure 8.6 is 40 and the experimental group data are marked with squares and the comparison group data are marked with circles. All scatter around the regression lines has been removed, so the true curvilinear regression surface is readily apparent. As revealed in Figure 8.6, given the curvilinear pattern in the data, there is no discontinuity in the regression surface at the cutoff score. A model that correctly fits the curvilinear shape to the data would correctly estimate that the treatment had no effect. However, incorrectly fitted (straight) regression lines have been added to the figure. Notice how the two straight regression lines exhibit a discontinuity at the cutoff score and differ in slope. In other words, fitting two straight regression lines rather than the correct curvilinear ones would make it appear as if the treatment caused a discontinuity in both level and slope, when in fact the treatment produced neither. In this way, fitting the wrong regression surfaces can result in biases in the estimates of the treatment effect.

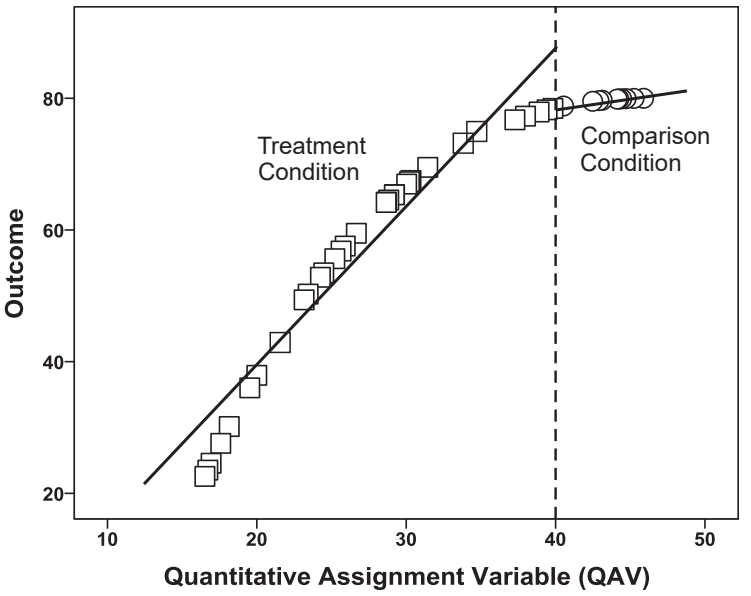


FIGURE 8.6. Hypothetical results from a regression discontinuity design showing how fitting straight regression lines in the presence of a curvilinear regression surface can lead to a bias in the estimate of a treatment effect.

The problem is there is no way to guarantee that a proper regression model will be fit to the data from an RD design. Suggestions for how to fit a correct regression model with which to estimate the effect of the treatment are presented next (Jacob, Zhu, Somers, & Bloom, 2012). One way to diminish the effects of misattributing a treatment effect because of misspecifying the regression model is to make the treatment effects as large as possible. (This recommendation also applies to any experimental design.) A large treatment effect is less plausibly explained as due solely to incorrectly modeled curvilinearity than is a small treatment effect.

8.3.1 Plots of the Data and Preliminary Analyses

The first step in the analysis is to create various plots of the data. One of the first plots should be one showing how the probability of receipt of the treatment condition varies according to the QAV. Perhaps the best way to create such a plot is to divide the data into thin slices according to the QAV scores on each side of the cutoff score. That is, the scores can be categorized according to intervals (bins) on the QAV. For example, if the QAV runs from 10 to 100 with a cutoff score at 40, the researcher could divide the QAV into categories of scores from 10 to 15, 16 to 20, 21 to 25, 26 to 30, 31 to 35, 36 to 40, and so on. Then the researcher plots the probability of receipt of the treatment condition within each bin. If everything has gone according to plan, the plot should look like the one shown in Figure 8.7, where a smooth line has been fit through the results from the bins. As in the figure, the plotted probabilities should be equal to one on the treatment side of the cutoff score and equal to zero on the comparison side. Such an

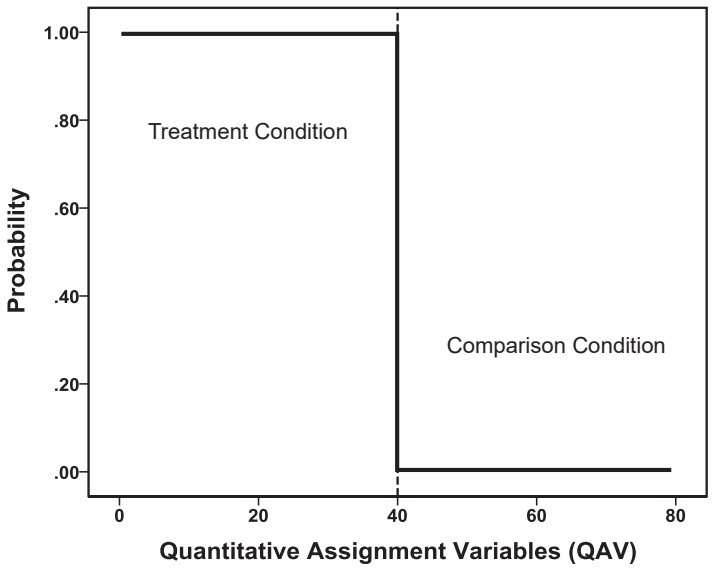


FIGURE 8.7. How the probability of receipt of the treatment condition varies with the QAV in a sharp regression discontinuity design.

outcome means the data come from a "sharp" RD design. This is in contrast to a "fuzzy" RD design. The present section (Section 8.3) assumes a **sharp RD design.** Section 8.4 considers the **fuzzy RD design.**

The second plot of the data should be a scatterplot as shown in Figures 8.1–8.4. The researcher should examine such a plot to see (1) once again, if participants have been assigned to treatments in accord with the cutoff score on the QAV, (2) if a discontinuity is visible at the cutoff score, and (3) to get a sense of the shape of the regression surface between the outcome variable and the QAV on both sides of the cutoff score. (Pay particular attention to outliers because they can influence the statistical analyses that follow.) The researcher should also look to see if discontinuities appear at other locations along the QAV besides at the cutoff score and perform the same types of analyses described below at such locations along the QAV. The presence of substantial discontinuities (i.e., pseudo treatment effects) at other locations besides the cutoff score without reasonable explanation for the presence of such discontinuities suggests that other unknown factors besides the treatment are operating, and so any discontinuity at the cutoff score might also be due to unknown factors besides the treatment (Imbens & Lemieux, 2008). Extra caution in interpreting a treatment effect estimate would then be warranted.

In addition to examining a scatterplot of the raw data, it can be useful to examine smoothed plots of the data. A smoothed plot can be produced using loess lines or splines that are fit separately to the data on each side of the cutoff score (Cleveland, 1993; Jacoby, 2000). Another option is to divide the data into thin slices along the QAV on each side of the cutoff score: that is, categorize the scores according to intervals (bins) on the QAV, as described above (Imbens & Lemieux, 2008). Then the researcher would calculate the mean of the outcome scores in each bin and plot those means on the vertical axis versus the midpoint of the bins on the horizontal (QAV) axis. The researcher can experiment with different bin widths until a plot is produced that provides an appropriately smooth regression surface. There are even computational methods for choosing an optimal bin width (Jacob et al., 2012; Lee & Lemieux, 2008). The smoothed plot would be examined (like the scatterplot) for apparent discontinuities (at the cutoff and elsewhere) and for assessment of the shape of the regression surfaces. The importance of plots is noted by Imbens and Lemieux (2008, p. 622), who emphasize that "if the basic plot does not show any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates with statistically and substantively significant magnitudes."

As will be explained shortly, different parametric models can be fit to the data. After each model is fit, plots of the residuals versus the QAV should be created and examined for deviations from what is expected if the regression surface between outcome and QAV was fit correctly. That is, the residuals should not deviate from a straight line when plotted against the QAV. Deviates from a straight line would indicate that the regression surface between outcome and QAV had been misfit. It is especially important to be sensitive to deviations near the cutoff score on the QAV because they would

indicate misfits of the regression surface that might produce spurious discontinuities. Again, loess regression or other methods can be used to smooth the data when examining plots of the residuals (West et al., 2014).

Curvilinearity in the regression of the outcome on the QAV can arise from various sources. A curvilinear regression surface could arise, for example, because of floor or ceiling effects. If curvilinearity is due to floor or ceiling effects, there would likely be a pile-up of scores at the ends of the frequency distributions of the outcome variables. Looking for such patterns in the data can help diagnose the likely presence of curvilinearity due to these sources.

Once researchers have a sense of the shape of the regression surface between the outcome scores and the QAV, they can use that knowledge to fit the correct regression surfaces to the data. Various methods can be used. If the regression surface is curvilinear, one approach is to apply a nonlinear transformation to the data to convert the curvilinear relationship into a linear one (Draper & Smith, 1998). But model-fitting approaches that use the untransformed data are more common. Three such approaches are presented next: global regression, local regression, and nonparametric regression. Nothing would prevent a researcher from comparing the results of more than one approach. Indeed, converging results from multiple methods increase researchers' confidence that conclusions are correct.

8.3.2 Global Regression

The standard approach to analysis of data from the RD design is to estimate the treatment effect using ANCOVA models. The task is to determine which ANCOVA model or models best fit the data (Trochim, 1984; Reichardt, Trochim, & Cappelleri, 1995).

8.3.2.1 Change in Level

In Figure 8.2, the regression lines of the outcome scores regressed onto the QAV are straight, the treatment produces a discontinuity at the cutoff point, and the treatment effect is constant at all values of the QAV. If the data from an RD design are structured as in Figure 8.2, the following model, which is the simplest ANCOVA model, will fit the data:

$$Y_i = \alpha + (\beta_T T_i) + (\beta_{\text{QAV}} \text{QAV}_i) + \varepsilon_i \tag{8.1}$$

where

Y_i is the score on the outcome variable for the ith participant;
T_i is an indicator variable representing the treatment assignment for the ith participant, where $T_i = 1$ if the participant is assigned to the treatment condition and $T_i = 0$ if the participant is assigned to the comparison condition;

QAV_i is the score on the QAV for the *i*th participant; and
ε_i is the error term or residual that represents all the factors not in the model and allows the data points, such as in Figures 8.1–8.4, to scatter around the regression lines.

This model fits parallel and straight regression lines on both sides of the cutoff score, as in Figures 8.1 and 8.2. Note that Equation 8.1 is the same as Equation 4.2 (see Section 4.6.1) and Equation 7.4 (see Section 7.4) except that the QAV_i variable has replaced X_i. In Equation 8.1, the estimate of α is the intercept of the regression line in the comparison group. This intercept is usually of little interest. The estimate of β_T is the treatment effect at the cutoff score. Because the regression lines are parallel, the estimate of β_T is equal to the vertical displacement of the regression lines at the cutoff score, as well as the vertical displacement everywhere else along the QAV. That is, if the treatment has an effect as in Figure 8.2, the treatment effect is constant across values of the QAV. The treatment does nothing more than shift the level of the regression line in the treatment condition vertically above or below the regression line in the comparison condition. The size of that shift is the estimate of β_T. The estimate of β_{QAV} is the slope of the two regression lines.

8.3.2.2 Change in Level and Slope

In Figures 8.2–8.4, the regression surfaces of the outcome scores regressed onto the QAV are straight lines, and the treatment produces a discontinuity, of one kind or another, at the cutoff score. But unlike in Figure 8.2, in Figures 8.3 and 8.4 the treatment effect varies across levels of the QAV: there is a change in slope due to the treatment. If the data from a RD design are structured as in Figure 8.3 or 8.4, the following model will fit the data, which allows for a change in both level and slope due to the treatment:

$$Y_i = \alpha + (\beta_T T_i) + (\beta_{QAV} QAV_i^*) + [\beta_{TQAV} (T_i \times QAV_i^*)] + \varepsilon_i \qquad (8.2)$$

There are two important differences to note between Equations 8.1 and 8.2. First, QAV_i in Equation 8.1 has been replaced by QAV_i^* in Equation 8.2, where QAV_i^* is equal to QAV_i minus the value of the cutoff score (i.e., $QAV_i^* = QAV_i - CS$, where *CS* is the cutoff score). This has the effect of shifting the placement of the *Y*-axis along the *X*-axis in the figures (by shifting the location of the zero point on the *X*-axis to the cutoff score). In Equation 8.1, the *Y*-axis is placed at the point where QAV_i equals 0. In Equation 8.2, the *Y*-axis is placed at the point where QAV_i^* equals zero, which is at the cutoff score. The second change is the addition of $[\beta_{TQAV} (T_i \times QAV_i^*)]$ to Equation 8.2. The notation $(T_i \times QAV_i^*)$ means that T_i is multiplied by (QAV_i^*) to create a new variable. Then this new variable is added to the statistical model multiplied by the corresponding coefficient β_{TQAV}. Adding this variable allows the slopes of the regression lines to differ across the treatment groups, as in Figures 8.3 and 8.4. The other variables in Equation 8.2 are as

defined in Equation 8.1. Note that Equation 8.2 is the same as Equation 4.4 in Section 4.6.2, except that QAV_i^* has replaced X_i^*.

With these two differences between Equations 8.1 and 8.2, the interpretations of the α and β coefficients also change. The coefficient α is still the intercept of the regression line in the comparison group, but that intercept has been moved because of the new placement of the *Y*-axis. The change in the location of the *Y*-axis makes the value of α equal to the height of the regression line at the cutoff score for the participants in the comparison condition. Because the regression lines are no longer parallel, the height displacement of the regression lines is not constant across the QAV variable. This means the treatment effect is not constant across the QAV variable. For example, the treatment effect is greater for smaller values of QAV than for larger values in Figure 8.3 and the reverse in Figure 8.4. By replacing QAV_i in Equation 8.1 with QAV_i^* in Equation 8.2, the estimate of β_T is the vertical displacement between the two regression lines at the cutoff score and so is the estimate of the treatment effect at the cutoff score. In other words, the treatment effect represented by β_T is estimated as the discontinuity in the regression lines at the cutoff score. If the cutoff score were not subtracted from the QAV_i score before the term was entered in the model as QAV_i^*, the discontinuity due to the treatment effect would be estimated where QAV_i (rather than QAV_i^*) equals zero, which is likely to be relatively uninformative, if not misleading.

Finally, the value of β_{QAV} is the slope of the regression line in the comparison group, and the value of β_{TQAV} is the difference in slopes between the regression lines in the treatment and comparison groups. As such, β_{TQAV} represents a treatment effect, which is the linear interaction of the treatment with the QAV variable. A test of the statistical significance of the estimate of β_{TQAV} is a test for the presence of that treatment effect interaction. When the estimate of β_{TQAV} is positive, the slope in the treatment group is greater than the slope in the comparison group, and vice versa. For example, if the slope in the comparison group is .5 and the slope in the treatment group is .75, the estimate of β_{TQAV} would be .25.

When the treatment effect varies (i.e., interacts) with the QAV, researchers might want to estimate the treatment effect at other points along the QAV (in place of or in addition to estimating the discontinuity at the cutoff score) to describe the nature of the treatment effect interaction. To estimate the treatment effect when the QAV equals some value QAV′, CS in the calculation of QAV_i^* should be replaced with QAV′. In Figure 8.4, for example, choosing a value for QAV′ less than the cutoff point would decrease the value of β_T because there is a smaller vertical discrepancy between the regression lines when the QAV equals QAV′ than when the QAV equals the cutoff score (CS), that is, at QAV*. In some circumstances, it is preferable to estimate the displacement of the regression lines in an RD design at a value other than at the cutoff score. Such circumstances arise, for example, if there were few participants in the experimental group so that the estimate of the regression slope in the experimental group would be unstable. In this case, it could be appropriate to estimate the displacement of the regression lines at a QAV score equal to the mean of the experimental group's QAV scores. This would make the value of β_T

equal to the average effect of the treatment in the experimental group, which would give the value of the average treatment effect on the treated (ATT) (see Cochran, 1957).

In general, however, methodologists suggest placing greater emphasis on differences at the cutoff score (CS) than anywhere else (West et al., 2014). The reason is that estimating a difference at the cutoff score requires less extrapolation of the regression lines than estimating a discontinuity elsewhere on the QAV. For example, estimating the displacement between the regression lines at a value of the QAV less than the cutoff in Figures 8.3 and 8.4 would mean extrapolating the regression line from the comparison group into a region where there are no scores from the comparison group. The further the extrapolation, the more uncertain is the estimate of the displacement.

A problem associated with fitting the interaction term in Equation 8.2 (as demonstrated in Figure 8.6) is that a curvilinear regression surface can be mistaken for an interaction. In other words, it might be difficult to distinguish between a curvilinear regression surface and an interaction of the treatment with the QAV. Consider Figure 8.6 again. If the data were scattered around the regression line (as would arise in practical circumstances), it might be difficult to choose between a curvilinear shape and a treatment effect interaction based on a visual inspection of the data. A statistical model with either (1) a regression surface with straight lines and a treatment effect interaction or (2) a curvilinear regression surface with no treatment effect (see below) could appear to fit the data quite well. The way to distinguish between the two specifications would be to test models where both interaction and polynomial terms were fit to the data (see below). Note that such problems are not as severe in a randomized experiment because there is no missing data on either side of a cutoff score (see Figure 8.5).

To reduce the potential for misinterpreting a curvilinear regression surface for a treatment effect interaction, some analysts suggest that the effect of the treatment be estimated only at the cutoff score and that claims about treatment effect interactions (i.e., claims about the β_{TQAV} parameter representing a treatment effect) should not be made (but see Lee, 2008; Jacob et al., 2012). Other analysts believe it is permissible to assess the effects of interactions but only when there is a statistically significant discontinuity at the cutoff score (e.g., Campbell, 1984). I do not share such views. I believe it is permissible to interpret an interaction, even when there is no discontinuity in level at the cutoff score. However, such interactions should be interpreted with great caution because a curvilinear relationship could be mistaken for an interaction. As noted earlier, estimates of differences in level at the cutoff score are generally more credible than estimates of differences at any other point along the QAV and generally more credible than the interpretation of the β_{TQAV} parameter as a treatment interaction (rather than as a result of a bias due to a curvilinear relationship).

8.3.2.3 Curvilinear Relationships

The statistical models presented so far assume that the regression surfaces in the two treatment groups are straight lines. As noted earlier, that need not be the case. The

traditional way to deal with curvilinearity in the regression surfaces is to fit ANCOVA models with polynomial terms in the QAV scores (Trochim, 1984). The following ANCOVA model is the same one as in Equation 8.2 but with both simple quadratic and quadratic interaction terms added:

$$\begin{aligned} Y_i &= \alpha + (\beta_T T_i) + (\beta_{QAV} QAV_i^*) + [\beta_{TQAV} (T_i \times QAV_i^*)] \\ &\quad + (\beta_{QAV2} QAV_i^{*2}) + [\beta_{TQAV2} (T_i \times QAV_i^{*2})] + \varepsilon_i \end{aligned} \tag{8.3}$$

where QAV_i^{*2} is QAV_i^* squared and the other variables are as specified for Equations 8.1 and 8.2. The estimate of β_{QAV2} reveals the degree of quadratic curvature in the data in the comparison group. The estimate of β_{TQAV2} reveals the difference between the treatment group and the comparison group in the degree of quadratic curvature. As before, the estimate of β_T is the discrepancy between the two regression surfaces at the cutoff score. The estimate of β_{QAV} is the slope of the tangent to the regression curve in the comparison condition, at the cutoff score (Cohen, Cohen, West, & Aiken, 2003). The estimate of β_{TQAV} is the difference in the slopes of the tangents to the regression curves in the treatment group compared to the comparison group, at the cutoff score. If the model is correct, the estimates of β_T, β_{TQAV}, and β_{TQAV2} all represent treatment effects. (Note that Equation 8.3 is the same as Equation 4.5 in Section 4.6.3, except that QAV_i^* has replaced X_i^*.)

Cubic terms can also be added to the model.

$$\begin{aligned} Y_i &= \alpha + (\beta_T T_i) + (\beta_{QAV} QAV_i^*) + [\beta_{TQAV} (T_i \times QAV_i^*)] \\ &\quad + (\beta_{QAV2} QAV_i^{*2}) + [\beta_{TQAV2} (T_i \times QAV_i^{*2})] \\ &\quad + (\beta_{QAV3} QAV_i^{*3}) + [\beta_{TQAV3} (T_i \times QAV_i^{*3})] + \varepsilon_i \end{aligned} \tag{8.4}$$

The interpretation of the parameters is a generalization of the interpretation of the parameters in Equation 8.3. Even higher-order polynomials can be added. In general, the order of the polynomial terms that needs to be added to fit a curvilinear shape is equal to one plus the number of inflection points in the regression surface. If there are zero inflection points, the order of the polynomial terms is one that means the straight-line model is used.

In theory, any curvilinear shape (and any interaction) can be fit with a model using a sufficient number of polynomial and interaction terms. But there is a limit to the number of terms that can be included in the model. I might also note that sentiment might be changing about the utility of cubic and higher-order models because the higher-order polynomial terms can produce overfitted regressions that can lead to dubious treatment effect estimates, especially with small datasets (Gelman & Imbens, 2018). As a result, it is argued, analysts should use local regression (which is described below) or nonparametric or semiparametric approaches such as splines instead of higher-order polynomial models. Even so, many analysts still list higher-order models as appropriate options (Angrist & Pischke, 2009; Bloom, 2012; Jacob et al., 2012; Wong et al., 2012).

8.3.2.4 Choosing among Global Regression Models

With regard to ANCOVA models that include polynomial and interaction terms, the task is to determine which terms should be included and which omitted to produce the best estimates of treatment effects. On the one hand, underfitting the model by including too few polynomial or interaction terms tends to bias the estimates of the treatment effects (and it can reduce power and precision). An example is provided in Figure 8.6 where fitting straight rather than curved regression lines produces spurious treatment effects. On the other hand, overfitting the model by including too many terms will not introduce bias. But overfitting the model by including too many terms can contribute to multicollinearity because the terms can be correlated among themselves and with the QAV. The problem with multicollinearity is that it can reduce power and precision and lead to unstable estimates of the treatment effect. An overfit model may lead to unbiased estimates of treatment effects but without sufficient power for those estimates to be statistically significant. Because overfit models avoid bias, the standard recommendation seems to be to err on the side of overfitting the model (Cappelleri & Trochim, 2015; Trochim, 1984; Wong et al., 2012). Because of the trade-offs, however, the researcher must navigate carefully between the two alternatives of underfitting and overfitting the correct model.

There are different strategies for determining the appropriate polynomial and interaction terms to add to the model to best fit the data. One strategy is to start with the most complex model that seems reasonable, given plots of the data or even a slightly more complex model than seems necessary. Then the researcher drops the highest-order terms one at a time if they are not statistically significant (Cappelleri & Trochim, 2015). For example, if the researcher started with an ANCOVA model with linear, quadratic, and cubic terms with interactions, the researcher would drop the cubic interaction term if it was not statistically significant. Then the model would be refit, and the cubic term would be dropped if it was not statistically significant. The model would yet again be refit and the quadratic interaction term would be dropped if it was not statistically significant, and so on. The reverse strategy is also possible: start with the simplest model and add statistically significant interaction and polynomial terms one at a time until no more terms are statistically significant. The problems with both of these model-fitting strategies are both power and multicollinearity. Because of multicollinearity among the terms, the statistical significance of polynomial and interaction terms is not necessarily a good indicator of whether these terms should be included in the model for bias reduction. A polynomial or interaction term could be statistically insignificant because of lack of power rather than because the term is not needed for a correct model fit.

When using ANCOVA models as just described, the following practice is recommended. Pay attention to indicators of multicollinearity such as the **variance inflation factor (VIF).** Equally important, see how treatment effect estimates vary across different models and examine residuals to help assess whether the regression surfaces have been adequately fit. Do not omit polynomial and interaction terms just because they are not statistically significant if the residuals indicate poor fit. If there is little change in

the size of the treatment effect estimate across a range of models and if there are good-looking residuals, use the simplest model (i.e., the model with the fewest polynomial and interaction terms) that is consistent with the data because it will provide the most precise treatment effect estimate.

An alternative strategy for selecting a polynomial model is to start with the simplest ANCOVA model and add polynomial and interaction terms based on indices of model fit (Jacob et al., 2012; Lee & Lemieux, 2010). Fit can be assessed by adding, to the ANCOVA model, indicator variables for the bins that were used to plot the data. If those indicator variables do not contribute significantly to the *R* squared of the simple ANCOVA model, the simple ANCOVA model is deemed adequate. If the indicator variables do contribute significantly to the *R* squared of the ANCOVA model, the next higher-order polynomial or interaction term is added, the test is conducted again, and so on. The logic of this procedure is the following. The indicator variables will fit any shape of regression surface and so will take up any lack of fit remaining after the ANCOVA model is fit. If the indicator variables do not significantly contribute to the *R* squared, the ANCOVA model can be said to fit the data sufficiently well without variance left over due to lack of fit. But again, the researcher should be cautious because a large sample size is required for the indicator variables to have sufficient power to indicate lack of statistically significant fit. Wing and Cook (2013) used yet a different strategy for selecting a polynomial model. For each model, they used least squares crossvalidation where they calculated the overall prediction error averaged across participants when each participant's scores were omitted from the model.

A researcher could also compare various ANCOVA models using goodness-of-fit statistics such as the Akaike information criterion (AIC) without using added indicator variables (Jacob et al., 2012). Such goodness-of-fit measures are sensitive to both the amount of residual variance in the model and the model's complexity (i.e., the number of parameters estimated in the model). For example, the AIC decreases as the residual variance decreases, and it increases as the complexity of the model increases. The AIC will be the smallest when the reduction in the residual variance is not offset by the increase in the model's complexity. The model with the smallest AIC is preferred. The problem is that goodness-of-fit measures such as the AIC can reveal which ANCOVA model fits best relative to other models, according to the given criterion, but not whether that best-fitting model provides a fit that is adequate in absolute terms (Jacob et al., 2012).

Once a best-fitting model is selected, sensitivity tests can be performed wherein the researcher deletes various percentages (say 1, 5, and 10%) of participants with the highest and lowest QAV scores (Jacob et al., 2012; van der Klaauw, 2002). The best-fitting model is then fit to each selected sample. To the extent that the estimates of the treatment effect remain the same as more and more data are deleted, the results are robust to model misspecification in the tails of the distributions. The researcher should pay attention to the size of treatment effect estimates more than to *p*-values in these sensitivity tests because dropping data will reduce power. In any case, researchers should

report the results of multiple analyses to reveal how estimates of treatment effects vary across models and datasets, and draw conclusions based on the range of estimates produced by the most plausible models (Jacob et al., 2012; Reichardt & Gollob, 1987). The most confidence should be placed in results that vary little across the range of plausible model specifications.

Additional covariates can be added to the ANCOVA model to increase power and precision (Imbens & Lemieux, 2008; Judd & Kenny, 1981; Schochet, 2008; Wing & Cook, 2013). Power and precision are maximized when the covariates are highly correlated with the outcome variable (above and beyond their relationships with the other variables in the ANCOVA model). Often, the most highly correlated measures are those that are operationally identical. This suggests that a pretest that is operationally identical to the posttest be collected and included in the model as a covariate, if it is not used as the QAV (Wing & Cook, 2013). However, adding covariates that are correlated with the QAV or the other variables in the model but not sufficiently correlated with the outcome can reduce power and precision because of multicollinearity.

8.3.3 Local Regression

An alternative to the global regression approach is local regression (Hahn, Todd, & van der Klaauw, 2001; Imbens & Lemieux, 2008; Jacob et al., 2012). In local regression approaches, data are deleted from the analysis except for the scores closest to the cutoff score, on each side. In other words, only the scores within a small distance (called the bandwidth) on each side of the cutoff score are included in the analysis. In local linear regression, the treatment effect is estimated by fitting the linear ANCOVA model (with an interaction term) to the selected data. By plotting the data, the researcher ensures that a linear regression surface appears to fit the data near the cutoff score. If not, polynomial terms might be added to the model. In any case, polynomial ANCOVA models can be used to provide sensitivity checks for the results of the linear model to see if the results stay the same across different models. Sensitivity checks could also be performed by varying the bandwidth. When narrowing the bandwidth, the number of needed polynomial terms should decrease (Angrist & Pischke, 2009). In addition, you can use the global approach as a sensitivity check for the local regression approach, and vice versa.

The difference between the global and the local regression approaches lies in how they trade-off bias versus precision. Unless the shape of the regression surface is fit perfectly, the estimate of the treatment effect in the global approach is bound to be biased and the bias tends to be greater with poorer model fits. To make matters worse, it is most important that the global model fits the data closest to the cutoff score, but the ANCOVA model does not give priority to those data. Indeed, the data farthest from the cutoff score (and especially outliers) can have an undue influence on the fit of the ANCOVA model at the cutoff score. The advantage of the global approach is that, if the ANCOVA model well fits all the data, including all the data in the analysis can increase

the precision of the treatment effect estimate. Conversely, the local regression approach reduces bias because it gives priority to data close to the cutoff score and because a simple linear model is likely to more closely fit the true regression surface the shorter is the interval of data over which the model is fit. But local regression suffers from reduced precision because it can omit a substantial portion of the data. The trick with the local regression approach is choosing an optimal bandwidth for the analysis that provides the best trade-off between bias and precision. Consult Jacob et al. (2012) for a discussion of two methods for choosing the optimal bandwidth, both of which are more complex than can be presented here. Also see Imbens and Lemieux (2008) and Thoemmes, Liao, and Jin (2017).

Because the local regression method throws away data, the choice between the global and the local regression methods depends substantially on the total sample size. The smaller is the total sample size, the less efficacious is the local regression method. But as sample size increases, the local regression approach becomes relatively more advantageous. Indeed, as the sample size increases and bandwidth decreases, bias is reduced and precision increases so that the trade-off between bias and precision becomes less important in the local regression approach. This is not the case with the global approach. Here bias remains the same as the sample size increases, even though precision increases. Large sample sizes favor the local regression approach over the global approach, especially when there is a relatively large amount of data near the cutoff score. But there are good reasons to investigate (and report) the results of all plausible analyses because confidence in the results will increase to the extent that the results agree (Imbens & Lemieux, 2008; Lee & Lemieux, 2010).

8.3.4 Other Approaches

The local regression approach is often characterized in the literature as a semiparametric or nonparametric procedure. Other nonparametric procedures for fitting the regression surface and estimating the treatment effect are also possible. These alternatives include Kernel and spline regression. These nonparametric methods are sufficiently complex that they warrant only the briefest of comment here. In **Kernel regression,** a regression surface is modeled by weighting scores nearby, more than scores far away from, the point at which the regression line is being estimated (Cohen et al., 2003). Different weighting schemes can be used. In a rectangular kernel weighting scheme, all the scores within a certain bandwidth are weighted equally, and scores outside the bandwidth are given no weight. But simple Kernel regression tends to be biased in the RD design (Bloom, 2012; Hahn et al., 2001; Lee & Lemieux, 2008). Darlington and Hayes (2017) explain how to use splines to fit nonparametric regression models to data (also see Shadish et al., 2011). Berk, Barnes, Ahlman, and Kurtz (2010) used the semiparametric approach of generalized additive models (GAMs) to fit data from an RD design (Shadish et al., 2011; Wong, Steiner, & Cook, 2013). Although alternative

nonparametric and semiparametric procedures are possible, the literature emphasizes the global and local regression approaches as described above, though that may be changing. As Wong et al. (2012, p. 325) note: "The current recommendations in the [regression discontinuity] literature is to employ multiple approaches to modeling the response function (parametric, nonparametric, and semi-parametric) and examine the extent to which the results present a consistent picture of program effects."

8.4 FUZZY REGRESSION DISCONTINUITY

As explained in Section 4.7, no-shows are participants assigned to the treatment condition who refuse to accept or otherwise fail to receive the treatment. Conversely, crossovers are participants who receive the treatment despite being assigned to the comparison condition. Just as in randomized experiments, some participants in an RD design might be no-shows and others might be crossovers. In an RD design, such noncompliance or nonadherence to treatment assignment results in what is called a "fuzzy" RD design (Campbell, 1969b; Trochim, 1984). The label arises because the assignment to treatments is no longer exclusively determined by the cutoff score on the QAV. Rather, there is some inexactness (i.e., fuzziness) in assignment because some participants with QAVs on one side of the cutoff score receive the treatment reserved for those with QAVs on the other side of the cutoff score. In practice, fuzziness appears to occur mostly near the cutoff score on the QAV. Compared to those with QAV scores far away from the cutoff score, participants with QAV scores just to one side of the cutoff are more likely to feel they should have fallen on the other side of the cutoff score and therefore should have been assigned to an alternative treatment condition. For the present section, I will assume the reason for both no-shows and crossovers is that the cutoff-based assignment to treatment conditions did not appropriately "nudge" participants into the assigned treatment condition. The condition where participants end up in the wrong treatment condition because of manipulation of the QAV of the cutoff score is addressed in Section 8.5.3.

The best approach, of course, is to try to avoid noncompliance with treatment assignment. In addition to participants switching treatment conditions, noncompliance sometimes arises because administrators who assign treatment conditions believe special circumstances accompany some participants who, as a consequence, are not assigned to treatments strictly on their QAV. If the reasons for such special dispensations can be quantified, the reasons for special dispensations can sometimes be used to create a composite QAV, which takes the extra criteria into account.

When noncompliance is suspected despite one's efforts to avoid it, noncompliance should be addressed in the analysis of the data. The first step in a fuzzy RD analysis is to plot (using bins on the QAV scores) the probability of treatment receipt on the vertical axis versus the QAV scores on the horizontal axis. In the sharp RD design, the plot

of probabilities will have the value of 1 on the treatment side of the cutoff score and the value of 0 on the comparison side of the cutoff score, as in Figure 8.7. In the fuzzy case, the probabilities will not be exclusively 0 or 1 but something less than 1 at places on the treatment side of the cutoff score and something greater than 0 at places on the comparison side of the cutoff scores, as in Figure 8.8. But there still needs to be a sharp discontinuity in the probabilities at the cutoff score. For example, there would be a discontinuity if the probabilities of receipt of the treatment condition were .8 and .2 on each side, right at the edge of the cutoff score, as in Figure 8.8. This is partial fuzziness. There would be no discontinuity, for example, if the probabilities of assignment were .5 and .5 on each side, right at the edge of the cutoff score. This would be complete fuzziness, and the analysis could not proceed as an RD design.

A participant who is a no-show can have a QAV score that is on only one side of the cutoff score. Similarly, a participant who is a crossover can have a QAV score that is only on the other side of the cutoff score. In this way, the distribution of no-shows and crossovers is discontinuous at the cutoff score. As a result, they can introduce spurious discontinuities in the regression surfaces. If fuzziness is limited to a narrow range around the QAV, so that there is no fuzziness beyond a narrow range, it is sometimes said that data in the narrow range might be omitted and the analysis performed as described in Section 8.3. Other solutions to the presence of no-shows and crossovers in the RD design are conceptually the same as those in the randomized experiment (see Section 4.7 and Sagarin et al., 2014). Next, solutions are briefly reviewed (Jacob et al., 2012).

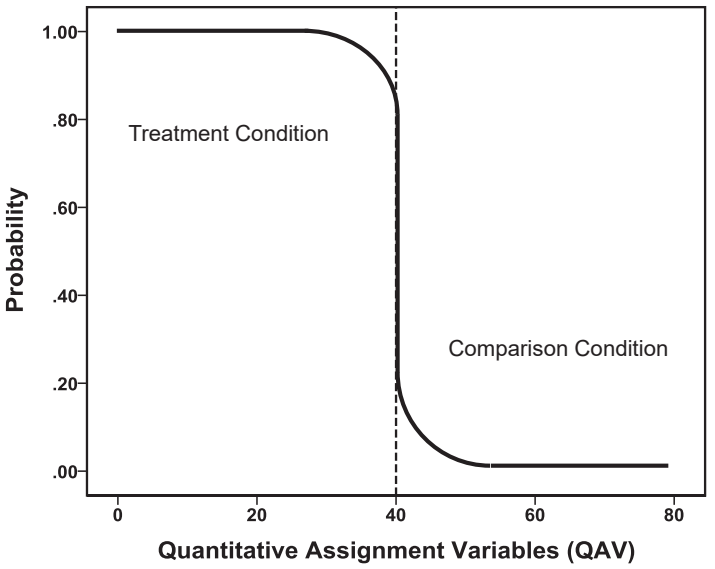


FIGURE 8.8. How the probability of receipt of the treatment condition can vary with the QAV in a fuzzy regression discontinuity design.

8.4.1 Intention-to-Treat Analysis

One approach is to ignore fuzziness and conduct the analysis as if fuzziness is not present. Those assigned to the treatment are compared to those assigned to the comparison condition, regardless of the treatment actually received. This is called the intention-to-treat (ITT) or treatment-as-assigned analysis. The effect that is estimated is the effect of the treatment on those offered the treatment whether or not they took advantage of it. As in the randomized experiment, the ITT analysis underestimates the effect of the treatment on those who complied with the treatment assignment. Just as with the randomized experiment, however, the direction of bias compared to the effect if there were full compliance is unknown. The ITT estimate may either over- or underestimate the effect of the treatment had there been perfect compliance to treatment assignment. For example, if on the one hand no-shows are those who would benefit most from the treatment, the ITT estimate can underestimate the treatment effect. On the other hand, the ITT estimate can overestimate the treatment effect if those who would have negative effects from the treatment are noncompliant. The ITT estimate can even have the opposite sign from the true treatment effect—negative when the treatment effect is positive and vice versa (Sagarin, West, Ratnikov, Homan, & Ritchie, 2014). The ITT analysis is sometimes the recommended approach when less than 5% of the participants are no-shows and/or crossovers combined (Judd & Kenny, 1981; Shadish & Cook, 2009; Trochim, 1984). In general, however, the current consensus seems to be to perform the CACE analysis that is described next.

8.4.2 Complier Average Causal Effect

The alternative to the ITT analysis is to estimate the complier average causal effect (CACE), which is also known as the local average treatment effect (LATE). The effect that is estimated is the effect of the treatment received at the cutoff score on those who complied with their treatment assignment (Foster & McLanahan, 1996; Hahn et al., 2001; van der Klaauw, 2002, 2008).

The assumptions of the CACE analysis for the RD design are the same as those for randomized experiments (see Section 4.7.4). The analysis assumes that participants who are no-shows receive the same treatment as the participants who are assigned to and receive the comparison treatment. The CACE analysis also assumes there are no defiers which is called the monotonicity restriction or assumption. It must also be assumed that no-shows would have the same outcome whether assigned to the treatment or comparison condition. The same holds for those who would cross over to receive the treatment (the always-takers). This is called the exclusion restriction.

The CACE analysis can be performed using two-stage least squares (2SLS) regression. Let D_i be an indicator variable indicating receipt of the treatment (where the value 1 indicates the participant received the treatment and the value 0 indicates the participant did not receive the treatment). Let T_i be an indicator variable indicating the

assignment of the treatment based on the QAV (where the value of 1 indicates the participant was assigned to the treatment condition and the value of 0 indicates the participant was assigned to the comparison condition).

In the first stage of the 2SLS analysis (assuming no change in slope due to the treatment), D_i is modeled as a function of T_i and the QAV score:

$$D_i = \alpha + (\beta_T T_i) + (\beta_{DQAV} \text{QAV}_i) + r_i \tag{8.5}$$

where QAV_i is the score on the quantitative assignment variable for the ith participant and r_i is the residual for the ith participant. This model is used to predict each participant's D_i score. Let $\hat{D}_i$ be the predicted, D_i, score for the ith participant.

In the second stage of the 2SLS analysis, the outcome scores (Y_i) are regressed onto the predicted, $\hat{D}_i$, scores and the QAV:

$$Y_i = \alpha + (\beta_D \hat{D}_i) + (\beta_{YQAV} \text{QAV}_i) + \varepsilon_i \tag{8.6}$$

where ε_i is the residual for the ith participant. The estimate of β_D is the CACE estimate, which is asymptotically unbiased (i.e., consistent) but biased in small samples.

Both equations are fit together using a 2SLS algorithm. If the equations were fit separately using ordinary least squares (OLS), the standard error from the second equation would need to be adjusted to obtain a correct value, given the uncertainty of estimating the values of D_i in the first stage.

Such an analysis can be conceptualized as an instrumental variable (IV) analysis where the T_i variable is used as the instrument for the D_i variable. The analysis assumes that the treatment assigned has no effect on the outcome except through its relationship with the treatment received. If the relationship between Y and the QAV is not linear, appropriate polynomial terms (of the same order) and interaction terms can be added to the models at both stages. Covariates can be added to both models to increase power and precision.

The local regression analysis for fuzzy regression is performed the same way as the global regression analysis (Hahn et al., 2001). The only difference is that, for the local regression analysis, the data would be restricted (in both equations) to the scores within the established bandwidth around the cutoff score. Jacob et al. (2012) show how to assess power in the fuzzy RD design.

8.5 THREATS TO INTERNAL VALIDITY

I have shown how fitting the wrong regression surfaces in the data analysis can bias estimates of treatment effects by introducing false discontinuities in both level and slope. That there should be no discontinuity in level or slope in the regression surfaces, except in the presence of a treatment effect, is called the **continuity restriction.** Other

threats to internal validity can also arise that violate the continuity restriction and thereby bias estimates of the treatment effects in an RD design. These other threats to validity include history effects, differential attrition, and manipulation of QAV scores.

8.5.1 History (Including Co-Occurring Treatments)

If another treatment or some other historical event arises in a way that corresponds to the cutoff score, it could bias the estimate of the treatment effect. Such events are sometimes called hidden treatments (Rubin, 2005). For example, consider a researcher interested in estimating the effects of Medicare on health outcomes. Medicare is made available to those who are 65 years of age and older, so age could be used as the QAV in an RD design. But other changes also occur at age 65, such as retiring from work and receiving the monetary benefits of senior citizen status, which may also have effects on physical health. Hence, an RD study of the effects of Medicare using age as the QAV would be assessing not just the effects of Medicare but the effects of all the other changes that might take place discontinuously at age 65. Researchers should try to avoid using a cutoff score that assigns participants to treatments other than the desired one, lest estimates of treatment effects be biased by co-occurring history effects.

8.5.2 Differential Attrition

Participants may differentially drop out from an RD study or fail to complete measurements in ways that co-occur with the cutoff score (Wong & Wing, 2016). For example, participants assigned to an undesired treatment condition may refuse to participate at all or choose to leave the study early in greater proportion than those assigned to the more desirable treatment condition. Such differential attrition would produce a compositional discontinuity, meaning that, concurrently with the treatment, the composition of the participants in the study changes in a discontinuous fashion at the cutoff point. As a result, differential attrition could cause a discontinuity in the regression surfaces at the cutoff score.

If compositional discontinuity due to attrition is present, a dropoff in the number of participants coinciding with the cutoff score would likely be present. Such a dropoff would be evident if the frequency distribution of the QAV scores were plotted (using either a histogram or a kernel density plot) and the distribution was not continuous at the cutoff score—showing a drop on the side that corresponds to the less desirable treatment. McCrary (2008) has provided a statistical test of the continuity of the QAV distribution at the cutoff score. The test involves using smoothed regressions to assess whether both sides of the frequency distribution meet at the cutoff score. The means of addressing attrition or other problems of missing data in an RD design are the same as the means of addressing attrition and missing data in a randomized experiment, as discussed in Section 4.8. As before, the best approach is to avoid attrition and missing data as much as possible.

8.5.3 Manipulation of the QAV

If the cutoff score is known ahead of time and QAV scores can be falsified or altered, either participants or administrators might be able to manipulate scores on the QAV to garner acceptance of an otherwise unqualified participant into a desired program. For example, Wong et al. (2012; Hallberg et al., 2013; Wong & Wing, 2016; also see Urquiola & Verhoogen, 2009) report an instance where some schools appeared to manipulate their scores to receive a rating of Adequate Yearly Progress under threat of negative consequences from No Child Left Behind legislation. Such manipulation could lead to a discontinuity in the relationship between the outcome variable and the QAV. Such manipulation can be avoided by keeping the cutoff score confidential and using QAVs that cannot be falsified or altered. In any case, it is always advisable to investigate the assignment process to determine if it appears to follow appropriate protocol.

Substantial manipulation of QAV scores would likely be evident in the frequency distribution of the QAV scores in the form of a localized bulge of scores on one side of the cutoff score and a corresponding deficiency on the other side. Again, the McCrary (2008) test could be useful in detecting such bulges or deficits.

8.6 SUPPLEMENTED DESIGNS

So far, I have considered only the basic RD design. This basic design can be supplemented, including in the five ways considered next.

8.6.1 Multiple Cutoff Scores

The basic RD design uses a single cutoff score to divide the participants into two treatment groups. An RD design with two or more cutoff scores could be used instead. For example, the treatment could be assigned to participants with scores in the middle of the QAV distribution, and the comparison condition could be assigned to participants at both ends of the QAV. In this way, there would be two cutoff scores. Or multiple cutoff scores could be used to assign participants to different doses of the treatment. The statistical models for the analysis of data from such designs would include additional indicator variables and interaction terms to assess discontinuities in level and slope at each cutoff score.

8.6.2 Pretreatment Measures

As noted in Section 8.3.2, pretreatment measures could be added as covariates to the analysis of data to increase power, though pretreatment measures could be used in other ways as well. Pretreatment measures could be used for falsification tests. The researcher would create the same types of plots and perform the same types of analyses

described in Sections 8.3 and 8.4 but using pretreatment measures in place of the real outcome variable (Imbens & Lemieux, 2008). That means the pretreatment measures would be regressed on the QAV to assess pseudo treatment effects. Because the pretreatment variables were collected before the treatment was introduced, there should be no discontinuities due to the treatment at the cutoff score. A discontinuity in a pretreatment distribution would raise suspicions that a similar discontinuity could be present in the real outcome variable at the cutoff score, even in the absence of a treatment effect. That is, a spurious discontinuity could mean the estimate of the treatment effect on the real outcome variable could be biased. Absence of a spurious discontinuity in the pretest measure, however, would strengthen the interpretation of the results derived using the real outcome measure.

When a pretreatment variable that is operationally identical to the outcome variable is available, analyses of the pretreatment variable can not only help assess the presence of spurious discontinuities but also help assess the shape of the regression surface in the posttest data in the absence of a treatment effect (Hallberg, Wing, Wong, & Cook, 2013; Wing & Cook, 2013). Because the treatment cannot have an effect on an operationally identical pretest measure collected before the treatment was implemented, any nonlinearity in the data must reflect true curvilinearity. To the extent that it is reasonable to assume the QAV would be related to the pretest and posttest measures in similar ways, the nature of curvilinearity in the pretest data can help specify the nature of curvilinearity in the posttest data. This procedure can also help estimate treatment effect interactions by comparing the pretest and posttest regression lines from the two scatterplots (where one scatterplot has the operationally identical pretest plotted on the vertical axis and the other scatterplot has the posttest plotted on the vertical axis). Parallel regression lines (from the two scatterplots) on the comparison side of the cutoff score accompanied by nonparallel regression lines (from the two scatterplots) on the treatment side of the cutoff score suggest a treatment effect interaction rather than a bias due to incorrectly fitting a curvilinear regression surface in the posttest data.

Wing and Cook (2013) go even further in the analysis of operationally identical pretest and posttest data (also see Angrist & Rokkanen, 2015). They suggest analyzing the pretest and posttest data together to create a pretest-supplemented RD design. Such an analysis would serve two purposes. First, the pretest-supplemented RD design with combined pretest and posttest data would increase the amount of data and thereby increase power and precision compared to an RD design with just the posttest data (i.e., a posttest-only RD design). Second, compared to a posttest-only RD design, a pretest-supplemented RD design would better enable researchers to extrapolate the posttest regression line in the comparison group into the region of the treatment group and thereby estimate effects at points on the QAV other than the cutoff score. In particular, Wing and Cook (2013) explain how to estimate the average treatment effect on all those who receive the treatment (the ATT) and not just the treatment effect at the cutoff score. However, the ATT estimate from the pretest-supplemented RD design would be credible

only to the extent that the two regression surfaces (one for the pretest and one for the posttest) on the comparison side of the cutoff score were parallel.

8.6.3 Nonequivalent Dependent Variables

If nonequivalent dependent variables are available, researchers should plot and perform analyses using these measures just as they would plot and analyze data from the real outcome variable. For example, Angrist and Pischke (2015) report an RD study where the treatment was expected to reduce traffic fatalities due to alcohol consumption (i.e., the real dependent variable) but not to influence other, non–alcohol-related causes of death (i.e., nonequivalent dependent variables). But the other, non–alcohol-related causes of death were likely to suffer from similar threats to internal validity as the alcohol-related causes of death. RD estimates were generated for both classes of outcome variables. That treatment effects were present for alcohol-related fatalities but pseudo treatment effects were not present for other causes of death added greatly to the credibility of the results.

8.6.4 Nonequivalent Groups

Instead of adding nonequivalent dependent variables, a researcher could attempt to obtain data on the QAV and outcome variable from a nonequivalent group of participants where the treatment was not implemented (Hallberg, Wing et al., 2013; Tang, Cook, & Kisbu-Sakarya, 2018; Wong et al., 2012). The same plots and analyses would be performed with the nonequivalent comparison group data as described in Section 8.3, as if a treatment had been implemented using the same outcome measure, QAV, and cutoff score. The nature of the relationship between the outcome measure and the QAV in the nonequivalent comparison group could again help diagnose the shape of the regression surface between QAV and outcome variables in the experimental data, in the absence of a treatment effect. That is, data from the nonequivalent comparison group would help researchers correctly model curvilinearities and treatment effect interactions in the data from the experimental group. In addition, the absence of pseudo treatment effects in the data from the nonequivalent comparison group could help rule out threats to internal validity.

Nonequivalent comparison groups could be created in two ways. First, data from the nonequivalent comparison group could come from a separate site where the data were collected contemporaneously with the data from the experimental site. The nonequivalent site would be selected to be as similar as possible to the experimental site, including sharing threats to internal validity. Alternatively, the nonequivalent comparison site could come from a cohort from a different time period. Such a design was implemented in assessing the effects of Medicaid on the number of visits to a physician where eligibility for Medicaid was determined using income as the QAV—with families earning less than $3000 per year being eligible for Medicaid (Lohr, 1972; Marcantonio

& Cook, 1994; Riecken et al., 1974). A measure of both the outcome variable (number of physician visits) and income (the QAV) was available from a nonequivalent cohort before Medicaid payments were first introduced in 1967. The relationship between the measure of physician visits and the QAV in the nonequivalent cohort (where these variables were collected before 1967) exhibited both no spurious treatment effect and a linear regression surface. This increased the credibility of fitting a linear model between the outcome variable and the QAV in the posttreatment cohort (where these variables were collected after Medicaid was introduced). This increased the researcher's confidence that the discontinuity observed in the posttreatment outcome data was due to Medicaid and not to misfit curvilinearity or threats to internal validity.

Tang et al. (2018) call an RD design that has been supplemented with a nonequivalent comparison group a comparative regression discontinuity (CRD-CG) design. They suggest analyzing the data from the nonequivalent comparison site and the experimental site together. The data from the untreated nonequivalent comparison site serves as an additional control but requires the stringent assumption that the regression surfaces in the comparison conditions in the two sites be strictly parallel. One benefit is that, assuming the required assumption is met, adding more participants in the form of an untreated nonequivalent comparison group can increase power and precision compared to a stand-alone RD design. As another benefit, the CRD-CG design could increase the confidence researchers have in extrapolating estimates of treatment effects beyond the cutoff score, as is also accomplished with the pretest-supplemented RD design (see Section 8.6.2).

8.6.5 Randomized Experiment Combinations

Yet another supplemented design would be an RD design combined with a randomized experiment (Aiken et al., 1998; Black, Galdo, and Smith, 2007; Boruch, 1975; Cappelleri & Trochim, 1994; Mandell, 2008; Moss, Yeaton, & Lloyd, 2014; Rubin, 1977; Trochim & Cappelleri, 1992). One such combined (design-within-design) approach would involve two cutoff scores: a lower cutoff score and a higher cutoff score. If the treatment were to be given to those most in need, those with a QAV score below the lower cutoff score would be assigned exclusively to the treatment condition (group A). Participants with QAV scores above the highest cutoff score would be assigned exclusively to the comparison condition (group D). Those participants with QAV scores in between the two cutoff scores would be assigned at random to either the treatment (group B) or the comparison condition (group C). The analysis would be undertaken in three parts. The comparison between groups B and C would be a randomized experiment. In addition, the researcher could perform two RD analyses (Rosenbaum, 1987). The first would compare groups A and B to group D using an RD comparison at the higher cutoff score. The second would compare groups C and D to group A using an RD comparison at the lower cutoff score. One advantage of such an elaborate design is that it would allow administrators to assign the treatment to those most in need while at the same time

recognizing that it would be fairest to give those with middling QAV scores an equal chance to be assigned to the treatment condition because QAV measures of need or merit are usually fallible. Such designs are said to contain a **tie-breaking randomized experiment.**

A different way to combine an RD design with a randomized experiment would be to add an RD component to one end of a randomized experiment. Consider a randomized experiment to compare an expensive treatment to a comparison condition consisting of no treatment but where it is inexpensive to collect the QAV and outcome measures. For example, the treatment might be inpatient care for substance abuse. Suppose, because of limited resources, the study can accommodate only 50 participants in the expensive treatment condition. Further, suppose 500 participants are available for the study but only the 100 neediest are to be enrolled in the randomized experiment. In such a study, 100 participants might be assigned to the randomized experiment, with the remaining 400 not included in the study at all. An alternative would be to add to the inexpensive comparison condition, using an RD assignment, the 400 who would otherwise be excluded from the study. That is, the 100 neediest would be enrolled in the randomized experiment where eligibility for the randomized experiment study would be based on a cutoff score on a QAV measure of need to create group A (those 50 in the randomized treatment condition) and group B (those 50 in the randomized comparison condition). Those on the other side of the cutoff score would form group C (the remaining 400) and would be measured on the QAV and outcome variable but, like group B, would not receive the treatment. Again, the analysis would be performed in parts. The first part would compare the 50 participants in group A to the 50 participants in group B in the form of a randomized experiment. The second part would compare the 50 participants in group A to the 400 participants in groups C in the form of an RD comparison. This makes use of the 400 participants who otherwise would have been discarded from the study. (A researcher could also compare group A to groups B and C combined.)

Designs that include both randomized experiments and RD designs as just described combine the strengths of both designs. The results of the combined designs can produce more credible results than designs that use just one design type or the other.

8.7 CLUSTER REGRESSION DISCONTINUITY DESIGNS

Section 4.9 described cluster-randomized experiments where random assignment to treatment conditions occurred at a higher (e.g., classroom) level, with data also being available at a lower (e.g., student) level. In a similar fashion, an RD design can have a cluster format (Pennel, Hade, Murray, & Rhoda, 2010). In this case, the RD assignment using a cutoff score would be made at the higher (e.g., classroom) level, with data also being available at the lower (e.g., student) level. That is, all the participants within a given cluster such as a classroom would be assigned to the same treatment condition,

with clusters being assigned to treatment conditions using a cutoff score on a cluster-level QAV. For example, Henry et al. (2010) estimated the effect of supplemental funding on student performance where schools were assigned to treatment conditions using, as a QAV, a measure of school-level educational advantage. The outcome measure was assessed at the school level, but data were also available for individual students.

As in the clustered-randomized experiment, the analysis of clustered RD designs can be performed using multilevel models. The model would be a simple extension of the models in the randomized experiment, with the QAV added as a covariate at the higher-level model. As in cluster-randomized experiments, power in RD discontinuity designs depends more on the number of clusters than on the number of participants within clusters (Schochet, 2008). (For a given number of participants, noncluster designs have more power and precision than cluster designs.) Adding covariates at both levels (but especially at the second level) of the analysis can greatly increase power (Schochet, 2009). For example, the analysis in Henry et al. (2010) included pretest measures of student achievement as covariates at the lower level of the analysis, as well as measures of school characteristics as covariates at the higher level of the analysis.

8.8 STRENGTHS AND WEAKNESSES

The RD design has advantages and disadvantages compared to both the randomized experiment and the nonequivalent group design. Consider relative advantages and disadvantages along the four dimensions of ease of implementation, generalizability of results, power and precision, and credibility of results.

8.8.1 Ease of Implementation

Randomized experiments require that those who receive the treatment condition are just as needy or meritorious as those who receive the comparison condition. When one treatment condition is perceived as more efficacious or desirable than another, withholding that treatment from equally needy or meritorious participants can be unpalatable to administrators, service providers, and participants alike. That is, stakeholders might resist studies in which equally needy or meritorious participants go unserved by a desirable treatment in a randomized experiment. In contrast, the RD design gives the presumed efficacious treatment to the neediest or most meritorious, leaving only the less needy or less meritorious in the comparison condition. In addition, an RD design (unlike a randomized experiment) can allow every eligible person in a population to be served—such as when assessing the effects of making the Dean's List on an entire population of undergraduate students. As a result, the RD design can sometimes be implemented in situations where the randomized experiment cannot be. Indeed, without any encouragement from researchers, administrators have sometimes assigned treatments based on a cutoff score on a QAV, so that the RD design

can sometimes be implemented after the fact using preexisting data in the form of a retrospective study. Such a possibility is rarely available with randomized experiments (Jacob et al., 2012).

Of course, if funding is provided to serve only *N* participants while the number of eligible needy or meritorious participants is 2*N,* a study cannot do anything more than serve half of those who are eligible. In that case, stakeholders can sometimes be convinced that random assignment is the fairest way to distribute the desirable treatment. But even if limited funding forces some needy or meritorious people to go unserved, the RD design allows the researcher to serve those who are *most* needy or meritorious among all those eligible for treatment. Such assignment might again be more appealing to stakeholders than random assignment, especially when criteria such as need or merit are well understood and perceived as appropriate.

Although it can be more palatable than random assignment, RD assignment is still demanding in that it requires participants be assigned to treatment conditions based strictly on a cutoff score on the QAV. In contrast, nonequivalent group designs require no such restrictions on assignment. Hence, nonequivalent group designs are generally easier to implement than RD designs (or than randomized experiments for that matter) and for that reason can be even more appealing to many stakeholders.

8.8.2 Generalizability of Results

Because randomized experiments can be so difficult to implement, the generalizability of results can be severely limited to those circumstances that permit random assignment. For example, participants willing to volunteer for a randomized experiment might be a much more circumscribed group than those willing to participate in an RD design. Hence, the generalization of results from a randomized experiment might be restricted to a more circumscribed population than the results from an RD design. On the other hand, estimates of treatment effects from an RD design are most credible when assessed at the cutoff score on the QAV, which can severely limit generalizability (though the pretest-supplemented RD design [see Section 8.6.2] and the comparative RD design with nonequivalent comparison group [see Section 8.6.4] design can help extend generalizability beyond the cutoff point). As compared to the results from the basic RD design, the results of a randomized experiment are not so restricted to treatment effect estimates near a cutoff score. In addition, researchers can assess how treatment effects vary across (interact with) different covariates in a randomized experiment. In contrast, researchers are limited in the RD design to assessing treatment effect interactions only with the QAV. But circumstances can arise where stakeholders are not interested in generalizing results in an RD design very far beyond effects near a cutoff score. For example, in a study of the effects of qualifying for the Dean's List (Seaver & Quarton, 1976), where the cutoff score was a GPA of 3.5, stakeholders would likely not be interested in generalizing results to students with GPAs of 2.0. That is, no one is likely to suggest that students with low rather than high GPAs be placed on the Dean's

List. Nor would stakeholders likely be interested in the effect of free or reduced-fee lunches on students with high, rather than low, family incomes.

Because nonequivalent group designs are most often even easier to implement than RD designs (and are not as focused on estimates of treatment effects at the cutoff score on a QAV), the results of nonequivalent group designs may be generalizable to an even broader population than are the results from RD designs.

8.8.3 Power and Precision

According to Goldberger's (1972a, 1972b, 2008) results, the RD design requires 2.75 times more participants than the randomized experiment to have the same precision, assuming the outcome scores and the scores on the QAV are multivariate normally distributed, equal numbers of participants are in each treatment condition, and the regression surfaces are linear. Cappelleri, Darlington, and Trochim (1994) extended those results to show, among other things, that the sample sizes needed for equal power and precision depend on the size of the treatment effect. Even with a large effect size, however, the RD design can still require 2.34 times more participants than the randomized experiment to have statistical power of .8 to detect the effect. If the regression surfaces are not linear, the differences in power and precision may be even greater (Jacob et al., 2012). An alternative way to describe the difference in precision is that the standard error for the treatment effect estimate in the regression discontinuity is generally between two and four times larger than the standard error in the randomized experiments when the two designs have the same sample sizes (Somers et al., 2013). Schochet (2009), along with Pennel et al. (2011), extended the results to clustered RD designs where the conclusions are similar to the results for nonclustered designs. Substantially larger sample sizes (generally, two to four times larger) are required in the clustered RD design to have the same power and precision as in clustered-randomized experiments.

The difference between the RD design and the randomized experiment in power and precision arises because of multicollinearity. In the randomized experiment, treatment assignment is uncorrelated with covariates in the statistical model. In contrast, treatment assignment in the RD design is highly correlated with the QAV. As noted above, adding interaction and polynomial terms to the RD model can worsen the degree of multicollinearity, further reducing power and precision (Jacob et al., 2012). In contrast, the randomized experiment can examine treatment effect interactions with greater power (and generality) than can the RD design because of reduced multicollinearity (though, even in randomized experiments, interactions are notoriously difficult to detect because of relatively low power). In addition (and as also noted below), larger sample sizes may be required, compared to the randomized experiment, to be able to assess and adequately fit a proper regression surface so as to avoid bias in the RD design. However, because the RD design can be easier to implement, RD designs may be able to enroll a larger number of participants negating some of the differences in power and precision that arise when the designs have equal numbers of participants

(Cook et al., 2008; but see Louie, Rhoads, & Mark, 2016). For example, as noted earlier, an RD design might be able to enroll an entire population, while a randomized experiment might be more limited to volunteers. In addition, the pretest-supplemented RD design (see Section 8.6.2) and the comparative RD design with nonequivalent comparison group (see Section 8.6.4) design can help increase the power and precision of an RD design.

The difference in power and precision between the RD design and the nonequivalent group design depends on the degree of overlap in the nonequivalent group design between treatment groups on the covariates in the model (which determines multicollinearity). In general, such overlap will be greater in the nonequivalent group design than in the RD design, and hence power may be greater in the nonequivalent group design than in the RD design.

8.8.4 Credibility of Results

The advantages in ease of implementation, generalizability, and both power and precision that the nonequivalent group design enjoys compared to the RD design are to a great extent negated by differences in the credibility of results. The results of RD designs tend to be more credible (in terms of the internal validity of causal inferences) than the results from nonequivalent group designs. In RD designs, selection into treatment conditions is determined by scores on the QAV; hence, the nature of selection differences is both known and measured. This is not the case in nonequivalent group designs where the causes of selection into treatment conditions are generally incompletely measured and may be largely unknown. As a result, it can be far more difficult to be confident that the effects of selection differences have been properly taken into account in the nonequivalent group design than in the RD design. This means researchers can generally have far less confidence in the estimates of treatment effects derived from nonequivalent group designs than from RD designs.

Such differences in confidence and credibility between the RD and nonequivalent group designs may seem counterintuitive. There is no overlap between treatment groups on the QAV in the RD design. In contrast, the nonequivalent group design is likely to have far greater overlap between the treatment groups on covariates. Intuitively, it might seem that treatment groups can better be compared, given overlap on covariates than given no overlap. But the lack of overlap on the QAV because assignment is based on a cutoff score is exactly what makes the RD design similar to the randomized experiment, as revealed in Figure 8.5. The nonequivalent group design does not enjoy the same similarity to the randomized experiment.

Although the RD design shares features with the randomized experiment, the results from randomized experiments are more credible than those from RD designs, at least in theory. The reason for the difference in credibility is that RD designs require that the relationship between the outcome and the QAV variable be properly modeled if a treatment effect is to be estimated without bias. The same is not required in a randomized

experiment when matching or blocking is used. As in the RD design, using ANCOVA in the randomized experiment requires fitting the correct regression surfaces. Because, however, there is no overlap on the QAV between the treatment groups in the RD design, fitting a proper regression surface in the RD design is more difficult than in the randomized experiment. In any case, fitting a proper regression surface requires a larger sample size in the RD design than in the randomized experiment because of the RD design's lack of overlap between the treatment groups on the QAV. It is such differences between the designs that make the results of well-implemented randomized experiments more credible than the results from even well-implemented RD designs, at least in theory.

The difference in credibility between the randomized experiment and the RD design also tends to increase as the researcher attempts to estimate the effect of the treatment, in the RD design, at locations along the QAV removed from the cutoff score. In the RD design, estimating the effect of the treatment at a location on the QAV other than at the cutoff score requires extrapolating a regression surface from one of the treatment groups into regions of the QAV where there are no data from that group. Randomized experiments do not suffer from the same limitation because the distributions of covariates overlap across the treatment groups.

As noted, the preceding differences in credibility between randomized experiments and RD designs are to be expected in theory. However, results in practice might be different. First, threats to internal validity such as due to attrition and noncompliance with treatment assignment can reduce the credibility of randomized experiments. So the difference in credibility between randomized experiments and RD designs may be diminished to the extent that attrition and noncompliance are less serious problems in the RD design than in the randomized experiment. In some instances, the appropriate comparison may be between a broken randomized experiment and a well-implemented RD design.

A substantial literature exists that empirically compares the results of RD designs to the results of randomized experiments. Studies in this genre include Aiken et al. (1998); Berk et al. (2010); Black et al. (2007); Buddelmeyer and Skoufias (2004); Green, Leong, Kern, Gerber, and Larimer (2009); Shadish et al. (2011); and Wing and Cook (2013). Also see the reviews of some of these studies by Cook et al. (2008) and by Cook and Wong (2008b). In most, but not all, of these studies, the results of the RD designs well mirror the results from randomized experiments. The conclusion to be drawn, based on such empirical considerations, is that carefully implemented RD designs can produce results that are about as credible as the results from carefully implemented randomized experiments (Hallberg et al., 2013; Shadish et al., 2011).

8.9 CONCLUSIONS

According to the What Works Clearinghouse, an RD design is eligible to receive the highest rating of "Meets WWC RDD Standards Without Reservations" (U.S. Department

of Education, 2017). But an RD design must be well implemented and the data well analyzed to satisfy the criteria necessary for that rating. Biases in the estimate of a treatment effect can arise in the analysis of data from RD designs if the regression surface (where outcomes are regressed onto the QAV) is not correctly modeled. Bias can also arise due to noncompliance, attrition, and the manipulation of the QAV. Strategies have been advanced for coping with at least some of these potential biases. The problem is that many of the strategies rely on sophisticated statistical methods that require strict, and often untestable, assumptions. The assumptions required for analyzing data from randomized experiments are often less strict than those required for analyzing data from RD designs, while the assumptions required for analyzing data from nonequivalent group designs are usually even stricter and often less plausible. The RD design can sometimes be implemented in research settings in which randomized experiments cannot be implemented. Also, the RD design can sometimes be implemented in place of a nonequivalent group design. To take advantage of the strengths of RD designs, researchers should be attuned to the conditions in which assignment to treatment conditions is made or can be made based on a cutoff score on a QAV.

A variety of statistical techniques have been advanced for properly estimating the regression surfaces in a RD design. The most common techniques include (1) estimating the complete regression surface by including polynomial terms in the regression models and (2) using linear models to estimate the regression surface only at scores close to the cutoff score on the QAV. The consequences of misspecifying the regression surface are not as serious in the randomized experiments and those consequences can be mitigated in a randomized experiment, by using matching/blocking instead of regression analysis in the form of ANCOVA. Neither matching nor blocking is an option in the RD design. In addition, the RD design requires a larger sample size than the randomized experiment to have the same power and precision.

The credibility of results from a carefully implemented RD design can approach that of the results from a carefully implemented randomized experiment. In addition, the RD design may produce more credible results than those produced by the randomized experiments if the randomized experiment suffers from attrition and noncompliance to treatment conditions when the RD design does not. The credibility of results from a carefully implemented RD design is generally greater than the credibility of results from even a carefully implemented nonequivalent group design.

8.10 SUGGESTED READING

Cook, T. D. (2008b). "Waiting for life to arrive": A history of the regression-discontinuity designs in psychology, statistics and economics. *Journal of Econometrics, 142,* 636–654.

—Presents a history of the RD design from its invention by Thistlewaite and Campbell (1960) to its recent resurgence among social scientists, especially economists.

The following articles, chapters, and books provide detailed and easy-to-follow descriptions of the RD design:

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect.* Princeton, NJ: Princeton University Press.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*(1), 43–82.

Hallberg, K., Wing, C., Wong, V., & Cook, T. D. (2013). Experimental design for causal inference: Clinical trials and regression discontinuity designs. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 1, pp. 223–236). New York: Oxford University Press.

More technical details are provided by:

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice, *Journal of Econometrics, 142*(2), 615–635.

Jacob, R., Zhu, P., Somers, M.-A., & Bloom, H. (2012). *A practical guide to regression discontinuity*. Washington, DC: Manpower Demonstration Research Corporation.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48,* 281–355.

9

Interrupted Time-Series Designs

This interrupted time-series design is one of the most effective and powerful of all quasi-experimental designs . . .

—SHADISH ET AL. (2002, p. 171)

Within a class of quantitative research designs, single-case experiments have a long and rich tradition of providing evidence about interventions applied both to solving a diverse range of human problems and to enriching the knowledge base established in many fields of science. . . .

—KRATOCHWILL, LEVIN, HORNER, AND SWOBODA (2014, p. 91)

Overview

In the basic interrupted time-series (ITS) design, multiple observations are collected both before and after a treatment is introduced. The observations before the treatment is introduced provide the counterfactual with which data after the treatment is introduced are compared. In particular, the effect of the treatment is estimated by projecting forward in time the trend in the observations before the treatment is introduced and comparing that projection to the actual trend in the observations after the treatment is introduced. Discrepancies between the projected trend and the actual trend are attributed to the effect of the treatment.

One of the relative advantages of ITS designs is that treatment effects can be estimated separately for individual participants. Such is not possible in between-groups comparisons (i.e., randomized experiments, nonequivalent group designs, and regression discontinuity designs) which can estimate only average treatment effects for groups of participants. In addition, ITS designs allow the temporal pattern of treatment effects to be assessed.

The ITS design can be implemented with a single unit or multiple units. Indeed, the design can be implemented with an entire population so that no participant goes without the treatment. The analysis of data from the ITS design depends on the number of time points and the number of units on which time series of data are collected. In general, the most troubling threats to internal validity in the ITS design are due to history,

instrumentation, selection differences, and misspecification of the trends in the time series of observations.

More complex ITS designs are also possible wherein multiple treatments are introduced across time or comparison time series are added. Such designs help rule out threats to internal validity such as those due to history, instrumentation, selection differences, and misspecification of trends.

9.1 INTRODUCTION

The basic interrupted time-series (ITS) design can be conceptualized as an extension of the pretest–posttest design, which was described in Chapter 6. A pretest–posttest design has a single pretest observation and a single posttest observation sandwiched around a treatment implementation. The design was diagrammed in Chapter 6 thusly:

$$O_1 \quad X \quad O_2$$

In this design, the treatment effect is estimated by comparing the pretest and posttest observations.

The basic ITS design adds multiple observations both before and after the treatment implementation, where both the pretest and posttest observations are strung out over time and all the observations are collected on the same variable. The design is diagrammed thusly:

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad O_5 \quad O_6 \quad X \quad O_7 \quad O_8 \quad O_9 \quad O_{10} \quad O_{11} \quad O_{12}$$

In this schematic, there are six pretest observations (O's) and six posttest observations spaced over time. The treatment (X) is implemented between the sixth and seventh observations. Depending on the nature of the study, a lesser or greater number of observations can be used, and the number of pretest observations need not be the same as the number of posttest observations. Indeed, an abbreviated ITS design could be implemented with an observation at a single posttreatment point in time. In the above diagram, that would mean the observations would stop at the seventh time point. But a single posttreatment time point would not allow the researcher to assess how the effect of a treatment can change over time (and would reduce the power and precision of the design). As will be explained below, one strength of the basic ITS design with multiple posttreatment observations is that it allows the researcher to assess how treatment effects change over time. In what follows, I will assume that observations are available at multiple posttreatment time points, as in the preceding diagram. The observations in an ITS design can be collected prospectively for the given research study. Or, as is often the case, the data could be obtained retroactively from an archive that records data assembled by others.

The data in an ITS design could be collected on a single person or separately on multiple people. For example, Nugent (2010) reports clinical studies in which individual persons were the units of analysis, including a study that assessed the effects of differential reinforcement to discourage wandering behavior in several individual persons with dementia, as well as a study of individual persons assessing the effects of hypnotic induction on panic attacks. Data could also be collected on a single conglomerate of persons (such as classrooms, schools, neighborhoods, hospitals, factories, countries, and so on) or separately on multiple conglomerates. For example, West, Hepworth, McCall, and Reich (1989) assessed the effects of drunk driving laws on traffic fatalities by collecting data from several entire cities. In general, data are collected on the same unit or units over time, but data could alternatively be collected from a different sample of units at each time point. For example, Smith, Gabriel, Schoot, and Padia (1976) assessed the effects of an Outward Bound program on self-esteem using an interrupted time-series design where a different random sample of participants was measured at each weekly time point.

The length of time (i.e., the lag) between observations can be anywhere from fractions of a second to years. The appropriate time lag depends on the research circumstances such as how quickly the treatment effect is expected to appear. For example, on the one hand, in studies of the effects of behavior modification on individual persons, the interval might reasonably be hours, days, or weeks. On the other hand, in studies of the effects of a change in a law or administrative rule (such as the effects on fatalities due to a change in speed limits or the enactment of laws requiring the wearing of seat belts), the interval between observations might reasonably be a month or a year.

As in the pretest–posttest design, the estimate of the treatment effect in an ITS design is obtained by comparing pretest observations to posttest observations, but the way in which the estimate is derived in the ITS design is more complex than that in the pretest–posttest design. In the ITS design, the trends over time in the pretest and posttest observations are each modeled statistically. The pretest trend is then projected forward in time and compared to the posttest trend. Treatment effects are estimated as the deviations between the projected and the actual posttest trends. To the extent that the treatment has an effect, the posttest trend should show a shift or interruption after the treatment is implemented, hence the name for the design. Of course, the observations at each time point must be taken on the same scale or measuring instrument for such estimates of the treatment effect to make sense. In addition, the pretest observations must be collected for a sufficiently long period of time for the researcher to be able to estimate the trend in the pretest observation and credibly project it forward in time. The number of posttest observations depends on the length of time over which a researcher wishes to estimate the treatment effect (or on limitations in the collection of data).

A hypothetical example of an ITS design is presented in Figure 9.1. The horizontal axis represents time, and the vertical axis represents the outcome variable. The jagged line shows how outcomes vary over time. The dashed vertical line in the figure at time point 20.5 marks the implementation of a treatment that took place between the 20th and 21st observations. The solid line that passes through the time series of observations

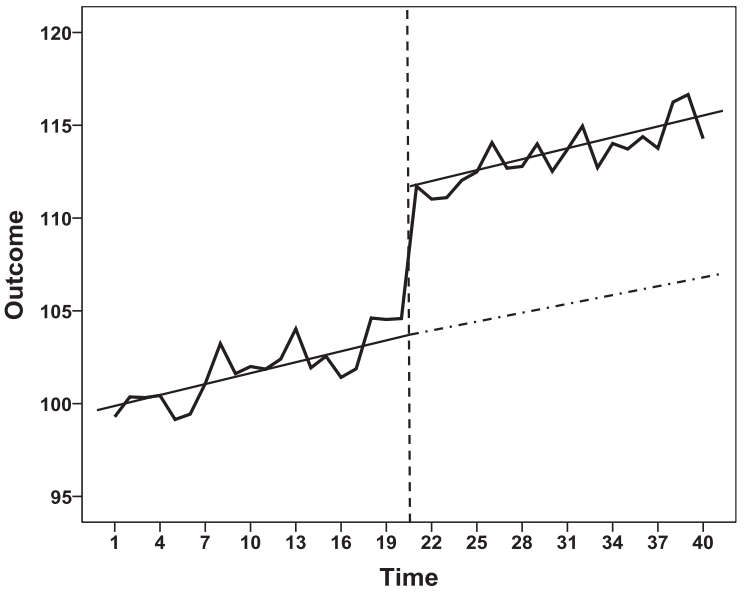


FIGURE 9.1. Hypothetical results from an interrupted time-series design where a treatment effect produces an abrupt change in level.

on the left-hand side of the figure reveals the trend in the pretreatment observations. This trend is upward. The solid line that passes through the time series of observations on the right-hand side of the figure reveals the trend in the posttreatment observations. The trend in the posttreatment observations is also upward and has the same slope as in pretreatment observations. The sloping dashed/dotted line is the projection of the pretreatment trend forward in time. The jump upward in the posttreatment trend line compared to the projected pretreatment trend line at the time the intervention is implemented is taken to mean that the treatment had an immediate positive effect. Such a jump is called a change in level. If the projected pretreatment trend line had coincided with the posttreatment trend line, the treatment would be estimated to have no effect.

The trends over time could be simpler than those shown in Figure 9.1. For example, in the Smith et al. (1976) study of the effects of Outward Bound on self-esteem, the pretreatment time series of observations was flat, showing neither upward nor downward slope. The treatment elevated the level of self-esteem, so that the trend line was higher after the treatment than before, but the trend after the treatment was also flat.

Another hypothetical example of an ITS design is presented in Figure 9.2. Again, time is represented on the horizontal axis, the vertical axis represents the outcome variable, the jagged line shows how outcomes varied over time, and the dashed vertical line in the figure at time point 20.5 marks the implementation of the treatment. As the solid line that passes through the time series of observations on the left-hand side of the figure reveals, the trend in the pretreatment observations is again upward. The solid line that passes through the posttreatment observations on the right-hand side of the

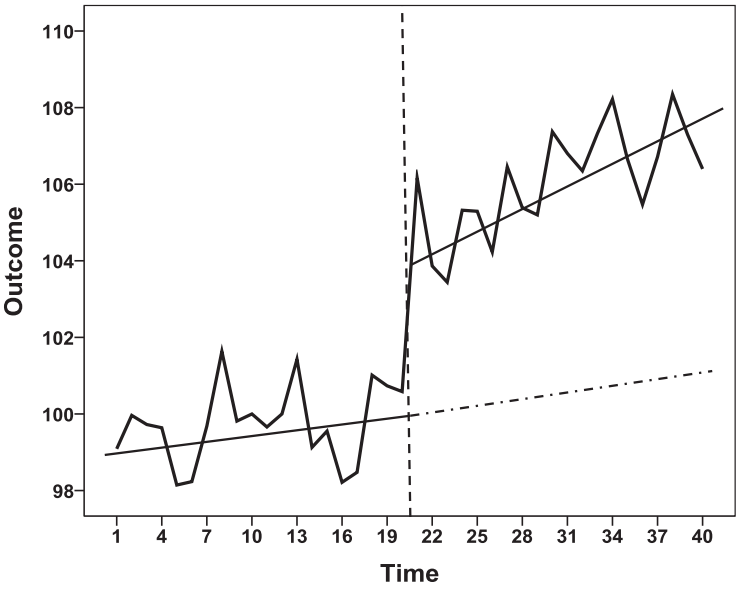


FIGURE 9.2. Hypothetical results from an interrupted time-series design where a treatment effect produces an abrupt change in level and slope.

figure is also upward but steeper than the pretreatment trend line. Again, the sloping dashed/dotted line is the projection of the pretreatment trend forward in time. As in Figure 9.1, there is a jump upward in the posttreatment trend compared to the projected pretreatment trend at the time the intervention is first implemented, indicating that the treatment had an immediate positive effect—change in level. That the slope of the posttreatment trend line is steeper than the projected pretreatment trend line is taken to mean that the effect of the treatment increases over time. Such a difference in the steepness of the trend lines is called a change in slope. More complex patterns of trends are also possible. For example, both the pretest and posttest trends could be nonlinear.

The ITS design is akin to the regression discontinuity design (Marcantonio & Cook, 1994). The RD design (see Chapter 8) assigns treatments to participants based on a cutoff score on a quantitative assignment variable (QAV). In the ITS design, the QAV is chronological time. That is, the ITS design assigns treatments using a QAV of chronological time based on a cutoff score, which is a given point in time. The ITS design is also similar to the RD design in that they are both often capable of providing highly credible estimates of treatment effects.

9.2 THE TEMPORAL PATTERN OF THE TREATMENT EFFECT

In Figure 9.2, the effect of the treatment is positive at the time of the intervention and is increasingly positive over time. Alternatively, the effect of the treatment could

be negative. For example, consider a weight-loss program assessed by taking weekly measures of a participant's weight both before and after a program is begun. In this case, the treatment might produce an immediate drop (rather than an increase) and a downward (rather than upward) slope in weight following the introduction of the program. A similar downward pattern occurred in a study by Steiner and Mark (1985), which assessed the effect of a community action group that mobilized in response to a bank's plans to raise mortgage rates for a group of consumers. The community action group called for a mass withdrawal of accounts from the bank. As a result, the bank's assets dropped immediately following the intervention. In addition, while there was a steep positive trend in assets before the intervention, that trend turned negative following the intervention. In other words, the negative effect of the intervention intensified over time.

As already noted, one strength of the basic ITS design is its ability to assess how the effect of a treatment can change over time. As also noted, a treatment effect can cause a change in level and/or a change in slope. The effect of the treatment might also be delayed rather than abrupt. A delayed effect in the form of a sleeper effect in research on persuasion is described in Cook et al. (1979). A delayed effect would also arise in a study of the effects of birth control on reproduction because of the nine-month gestation period, if data were collected at monthly intervals. An effect could also be delayed because the effects of a program diffuse slowly, are introduced gradually, or because a buildup to a certain threshold level is required before an effect is evident. The effect of a treatment could also change over time in a nonlinear fashion (such as when an effect is transitory or temporary rather than permanent). Figure 9.3 graphically presents a few ways in which a treatment effect might be manifest. Although not displayed in Figure 9.3, a treatment could also change the variability of the posttreatment observations rather than the mean level of the time series, although such a change is relatively rare.

In the basic ITS design, the further the pretreatment trend must be extrapolated into the future (so as to create a counterfactual for comparison with the observed posttreatment trend), the less certain that projection becomes. As a result, researchers can generally be most confident of a treatment effect estimate at the precise point in time when the treatment is introduced and less confident of an estimate farther removed from the point in time when the treatment is introduced. For example, researchers can generally be more confident that an abrupt change in level in an ITS design is due to the treatment than that a change in slope (i.e., a gradual effect) is due to the treatment because assessing a change in slope means the pretest trend must be extrapolated beyond the point in time at which the treatment is introduced.

With gradual or delayed treatment effects, confidence in the estimates of the treatment effects can sometimes be increased by comparing the obtained pattern in the ITS data to the pattern of effects predicted to arise based on additional data. For example, Hennigan et al. (1982) used data on the proportion of households owning televisions to predict the pattern of expected outcomes in assessing the historical effects of TV viewing on crime. The proportion of households owning televisions increased gradually

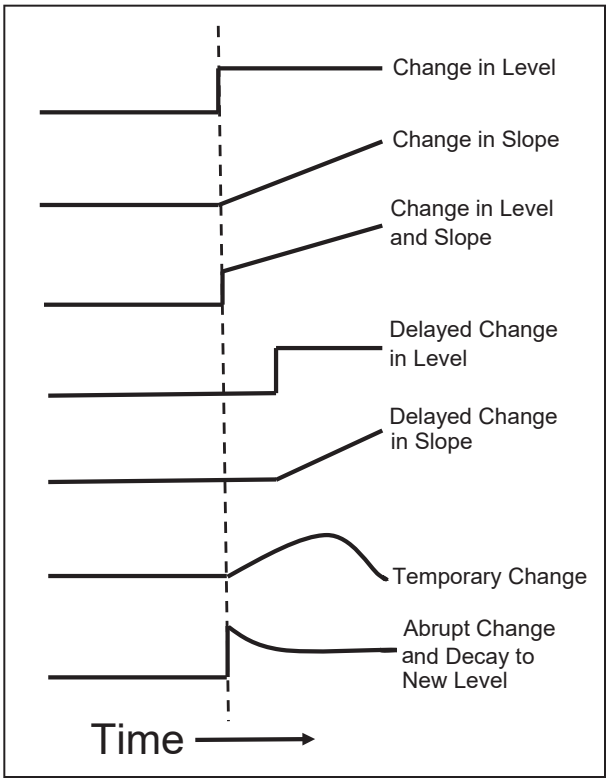


FIGURE 9.3. Different patterns of treatment effects over time. The dashed line indicates when the treatment was implemented. The solid lines show the trend in observations both before and after the intervention.

over time, so Hennigan et al. (1982) expected a correspondingly gradual effect of TV viewing on crime. The benefits of incorporating additional data into an ITS design are discussed further in Section 9.8. In addition, uncertainty about estimates of treatment effects that are delayed is reduced to the extent that the length of the delay can be predicted a priori. When the expected lag in a treatment effect is unknown, delayed treatment effects increase the chances that something other than the treatment (such as history effects: see Section 9.6) causes observed changes in the time series.

9.3 TWO VERSIONS OF THE DESIGN

As already noted, data in an ITS design could be collected from a single unit or from multiple units where the units could be either individual people or aggregates of people. The letter *N* will be used to denote the number of units in the ITS design. If on one hand *N* is 1, a single time series of observations is collected on a single unit, whether that unit is an individual person or an aggregate of persons such as a classroom, school, or

community. For example, McSweeny's (1978) study collected data on telephone usage from the city of Cincinnati to produce a single treatment time series of observations. On the other hand, if N is 30, multiple time series of observations are collected—one time series of observations for each of 30 different units (where, again, a unit might be either an individual person or an aggregate of persons such as a classroom, school, or community). For example, Nugent, Champlin, and Wiinimaki (1997) assessed the effects of anger control training on delinquent behavior on each person in a sample of 109 adolescents. Somers et al. (2013) assessed the effects of the Reading First program (which is a program of reading instruction) on a sample of schools where the individual school ($N = 680$) was the unit of analysis. In addition, let J be the number of time points in a design. For example, if there are 10 pretreatment observations spaced over time and 10 posttreatment observations spaced over time, $J = 20$.

The analysis of data from an ITS design depends on the values of N and J. Obviously, many combinations of N and J are possible, and not all possibilities can be addressed here. To make the presentation manageable, I have divided the discussion into two primary sections based on the size of N. The first section is for ITS designs when $N = 1$ (or N is small, and each time series is analyzed individually). The second section is for ITS designs where N is large (perhaps of size 30 or more) and the N individual time series of data are analyzed together as a group of individual time series.

Regardless of the sizes of N and J, the first step in the analysis of data from an ITS design, as in all designs, is to plot the data (as in Figures 9.1 and 9.2). A plot can reveal the likely shape of a treatment effect. Examining plots of the data also helps determine the likely equations that will need to be fit to the data to properly model trends over time. Plots of data can also help reveal if analytical assumptions are correct and the type of adjustments necessary if the assumptions appear violated. Incorrectly modeling a curvilinear trend in the data by fitting a linear trend can bias a treatment effect estimate in the ITS design, just as it can in the RD design (see Section 8.3). Only by examining the data and fitting proper models can researchers hope to avoid such biases. It is all too easy to fit a model to the data and arrive at incorrect conclusions because the model does not fit the data as the researcher presumes. The analyst needs to examine the data (including estimated residuals) to make sure the models fit properly so that the interpretations of the results are correct.

9.4 THE STATISTICAL ANALYSIS OF DATA WHEN $N = 1$

As noted, to estimate the treatment effect, the researcher models the trends in both the pretreatment and posttreatment time-series of observations. Then the trend in the pretreatment time series data is projected forward in time and compared to the posttreatment trend in the data. As also just noted, if the trends in the data are curvilinear but are modeled as straight lines, the projection of that pretreatment trend and comparison with the posttreatment trend can lead to biased estimates of the treatment effect. So,

correctly modeling the trends in the data is a central task in the analysis of data from ITS designs.

The present section concerns the proper modeling of ITS data when $N = 1$. When $N = 1$, there is a single time series of observations as in Figures 9.1 and 9.2. The present section is divided into two subsections based on the size of J. The first subsection considers the case where $N = 1$ and there are observations at many time points (i.e., J is large, preferably 50 to 100 or more). The second subsection considers the case where $N = 1$ and there are observations at relatively few time points (i.e., J is as small as 8 to 10).

9.4.1 The Number of Time Points (*J*) Is Large

If the trends in the observations are linear and the treatment has an abrupt effect in both level and slope (as in Figure 9.2), the following model could be fit to the data:

$$Y_j = \alpha + (\beta_1 \text{ POSTINTERVENTION}_j) + (\beta_2 \text{ TIME}_j^*) + [\beta_3 (\text{POSTINTERVENTION}_j \times \text{TIME}_j^*)] + \varepsilon_j \quad (9.1)$$

where

j is a subscript representing time;
Y_j is the outcome variable at time j, so that Y_1 is the first observation, Y_2 is the second observation, and so on;
$\text{POSTINTERVENTION}_j$ is an indicator variable that equals 0 before the intervention is introduced and 1 afterward;
TIME_j is a variable representing time: coded 1 for the first observation in time, 2 for the second observation in time, and so on (assuming the observations are equally spaced);
TIME_j^* is (TIME_j minus TIMEI) where TIMEI is the time of the intervention;
$\text{POSTINTERVENTION}_j \times \text{TIME}_j^*$ is the product of $\text{POSTINTERVENTION}_j$ and TIME_j^*;
ε_j is the residual, or how much each individual observation deviates from the trends in the pretreatment and posttreatment data;
α is the level of the pretreatment time series at TIMEI;
β_1 is the change in the level of data from before to after the intervention at TIMEI (based on a comparison of the pretreatment and posttreatment trends);
β_2 is the slope of the observations before the intervention; and
β_3 is the change in the slope from before to after the intervention.

In Equation 9.1, the value of β_1 is the change in level that is attributed to the effect of the treatment. The value of β_3 reveals how much the treatment effect changes the slope of the observations. For example, if the slope of the observations is .5 before the intervention and 1.25 afterward, then β_3 is equal to the difference between .5 and 1.25,

which is .75. In Equation 9.1, TIMEI must be subtracted from $TIME_j$ to create the $TIME_j^*$ variable, so that the change in level and the change in slope are estimated at the point in time at which the intervention begins (Huitema & McKean, 2000; Rindskopf & Ferron, 2014).

If there is a linear trend in the data over time and if the treatment abruptly alters the level of the time series of observations without altering the slope (as in Figure 9.1), the following model would fit the data:

$$Y_j = \alpha + (\beta_1 \text{ POSTINTERVENTION}_j) + (\beta_2 \text{ TIME}_j^*) + \varepsilon_j \quad (9.2)$$

The difference between Equations 9.1 and 9.2 is that the interaction term ($\text{POSTINTERVENTION}_j \times \text{TIME}_j^*$) has been dropped. In Equation 9.2, the value of β_1 is the treatment effect, which is simply a change in the level of the observations at the time the intervention is introduced. I present Equation 9.2 mostly for comparison with Equation 9.1. In general, fit Equation 9.1 rather than fitting Equation 9.2 so that the presence of a change in slope can be assessed empirically—by examining the estimate of the β_3 parameter in Equation 9.1.

If the treatment effect is more complicated than a change in slope and level, the $\text{POSTINTERVENTION}_j$ variable can be replaced, for example, with a variable or variables that have a more complex pattern than 0's followed by 1's (which is called a step function). For example, if the treatment had an abrupt but temporary effect, the $\text{POSTINTERVENTION}_j$ variable could be replaced with a variable that, over time, takes on the values of 0, then 1, and then back to 0 again. In addition, if the trends in the observations were curvilinear rather than straight lines, polynomial terms in $TIME_j^*$ could be added to the model to mirror the curvilinearity. The disadvantage of adding higher-order polynomial terms is that the resulting models can overfit the data, leading to erratic estimates of treatment effects. There are also alternatives to using the polynomial approach to estimate curvilinear trends. The alternatives include spline approaches such as generalized additive models (Fox, 2000; Shadish, Zuur, & Sullivan, 2014; Simonoff, 1996; Sullivan, Shadish, & Steiner, 2015). Such techniques are more complex and require more advanced software than the polynomial procedures. Because of the complexities involved with the advanced methods, polynomial models seem to be the most popular options in the social and behavioral sciences literatures.

Equations 9.1 and 9.2 can be fit to data using ordinary least squares (OLS) regression but not without potential problems. OLS regression assumes that the residuals (the ε_t's) are independently distributed, which is not typically the case in time-series data. Instead of being independent, the residuals are often **autocorrelated** (also called serially dependent), which means they are related to one another across time periods. Typically, when residuals are autocorrelated, residuals closer together in time are more highly correlated than residuals further separated in time. For example, residuals at times j and $j + 1$ are likely to be more highly correlated than residuals at times j and $j + 2$. Residuals can also display **cyclical autocorrelation** (also called seasonality) because

residuals at similar cyclical time points—say, 12 months apart—are correlated. Cyclical autocorrelations can arise, for example, because of month-to-month patterns across years, day-to-day patterns across weeks, or hour-to-hour patterns across days.

In the presence of autocorrelated residuals, OLS regression produces unbiased estimates of treatment effects but biased estimates of their standard errors. On one hand, when the autocorrelations are positive, the estimated standard errors are too small, making confidence intervals too narrow and statistical significance tests too liberal (i.e., too likely to reject the null hypothesis). Temperatures throughout a day, for example, tend to be positively autocorrelated. High temperatures at one time during the day tend to predict relatively hot temperatures at a later time during the day. The same holds for cold temperatures. On the other hand, when autocorrelations are negative, the reverse holds. That is, the estimated standard errors are too large, making confidence intervals too wide and statistical significance tests too stringent. The amount eaten during a meal might produce negative autocorrelations. Eating a large meal for lunch might mean a person tends to eat relatively less at dinner. Conversely, eating little for lunch might mean a person tends to eat relatively more at dinner. Although negative autocorrelations are possible, positive autocorrelations are more common.

The point is that, in the presence of substantial autocorrelation, OLS regression must be modified or replaced by procedures that take account of the autocorrelation in the data (including patterns of autocorrelation that are cyclical). The most common way to take account of autocorrelation is with the use of **autoregressive moving average (ARMA) models** (Box, Jenkins, & Reinsel, 2008; McCleary, McDowall, & Bartos, 2017). As the name implies, these models allow for two types of autocorrelation: autoregressive and moving average autocorrelation. Consider each in turn.

Autocorrelation in the observed outcome scores (Y) is modeled as autocorrelation in the residuals (ε) in Equation 9.1 or 9.2. In an **autoregressive (AR) model,** current residuals are dependent on past residuals. For example, if ε_j is the residual at time j, a first-order autoregressive model (AR(1)) for the residuals would be

$$\varepsilon_j = (\varphi_1\ \varepsilon_{j-1}) + u_j \tag{9.3}$$

where ε_{j-1} is the residual from the immediately prior time period, φ_1 is the first-order autoregressive coefficient with a value between –1 and 1, and u_j is a random (uncorrelated over time—also called **white noise**) error. Because the current residual (ε_j) is composed of the immediately past residual (ε_{j-1}), current and past residuals are correlated. With autoregressive models, a current residual is affected by all past residuals because ε_j is influenced by ε_{j-1}, which is influenced in turn by ε_{j-2}, which is influenced in turn by ε_{j-3}, and so on. In other words, the residuals have a never-ending memory, though that memory tends to fade over time because φ_1 is usually less than 1 in absolute value. In particular, the effect of ε_{j-1} on ε_j is φ_1, while the effect of ε_{j-2} on ε_j is φ_1^2, the effect of ε_{j-3} on ε_j is φ_1^3, and so on. If you hold grudges and the strength of your grudges decays

but never dissipates completely over time, your grudges might well be modeled with an autoregressive process.

A second-order autoregressive model (AR(2)) would be

$$\varepsilon_j = (\varphi_2\ \varepsilon_{j-2}) + (\varphi_1\ \varepsilon_{j-1}) + u_j \tag{9.4}$$

where ε_{j-2} is the residual from two prior time periods. Higher-order AR models are similarly defined. AR(p) means the AR model is of order p, where current residuals are composed of the p prior residuals. The AR(2) model is like the AR(1) model in that the effect of a residual reverberates forever over time, though with diminishing strength. The difference is that the AR(2) model allows for even greater reverberation over time than the AR(1) model.

Like an autoregressive model, a **moving average (MA) model** exhibits a memory but in a different manner. A first-order moving average (MA(1)) model is

$$\varepsilon_j = (\theta_1\ u_{j-1}) + u_j \tag{9.5}$$

where u_j is a random (white noise) error, u_{j-1} is the error from the prior residual and θ_1 is the moving average coefficient with a value between –1 and 1. In such models, the current residual (ε_j) shares an error term (u_{j-1}) with the prior residual. Because of the shared error term, the residuals are correlated. An MA(1) residual is influenced by the error term from the immediately preceding past residual, but the influence stops there. The error term from two prior time periods has no influence on the residual now. It is as if your grudge from yesterday is remembered today but forgotten by tomorrow.

A second-order moving average (MA(2)) model is

$$\varepsilon_j = (\theta_2\ u_{j-2}) + (\theta_1\ u_{j-1}) + u_j \tag{9.6}$$

where u_{j-2} is the error term for the residual from two prior time periods. An MA(2) model exhibits more memory than an MA(1) model, but both limit the number of time periods over which an autocorrelation is present. To return to the example just given, with an MA(2) model, your grudge would be held for two days before it was forgotten. Higher-order moving average models are defined similarly. MA(q) means the MA model is of order q, where current residuals are composed of q prior error terms.

It is also possible to have combined autoregressive and moving average (ARMA) models that are simply combinations of the models given above. West and Hepworth (1991) provide path diagrams of ARMA models that help make the differences clear. The presence of autocorrelation due to cyclical effects introduces even greater complexity into the models (McCleary et al., 2017).

Fitting a model with an ARMA structure for the error terms usually proceeds in three steps (Box et al., 2008; McCleary et al., 2017). The purpose of the steps is to

diagnose the nature of serial dependency in the data so that it can be modeled and its influence removed.

The first of the three steps is called the identification phase. A model is fit to the data (such as Equations 9.1 or 9.2) using OLS regression, and the residuals from the model are saved. The residuals are then examined to identify the type of AR, MA, or ARMA model that best accounts for the autocorrelations among the residuals. Toward this end, an **autocorrelation function (ACF)** and a **partial autocorrelation function (PACF)** are calculated for the residuals. The first entry in the autocorrelation function (ACF_1) is the correlation between the residuals and themselves after they have been lagged one time period. That is, the first entry in the autocorrelation function is a correlation calculated by pairing ε_j with ε_{j-1}, ε_{j-1} with ε_{j-2}, ε_{j-2} with ε_{j-3}, and so on. The second entry in the autocorrelation function (ACF_2) is the correlation between the residuals and the lag-two residuals. The third entry in the autocorrelation function (ACF_3) is the correlation between the residuals and the lag-three residuals, and so on. For example, suppose the time series of residuals is 6, 3, 7, 4, 2, 8. . . . Then the residuals are lined up thusly:

Time period:	1	2	3	4	5	6 . . .		
Residuals:	6	3	7	4	2	8 . . .		
Lag-one residuals:		6	3	7	4	2	8 . . .	
Lag-two residuals:			6	3	7	4	2	8 . . .

Correlations are calculated based on how the data are lined up. For example, the ACF_1 is the correlation between the residuals and the lag-one residuals, as lined up above. The ACF_2 is the correlation between the residuals and the lag-two residuals, as lined up above. (Note that some of the residuals are not included in the computation of the correlations because they are not paired with lagged residuals.)

The partial autocorrelation function is the same as the autocorrelation function except partial correlations are calculated rather than correlations, holding intermediate residuals constant. The first entries in the ACF and the PACF are the same. The second entry in the partial autocorrelation function ($PACF_2$) is the partial correlation calculated by pairing the residuals with themselves lagged two time periods, while holding constant the residuals lagged one time period. The third entry in the partial autocorrelation function ($PACF_3$) is the partial correlation calculated by pairing the residuals with themselves lagged three time periods, while holding constant the residuals lagged one and two time periods, and so on.

AR(p) and MA(q) models produce different patterns in the ACF and PACF. An AR(1) model has nonzero expected values for all ACF values, but the size of these values diminishes exponentially, with the time lag based on the value of φ_1. For example, the ACF for an AR(1) model has expected values as in Panel A of Figure 9.4. For a φ_1 value of .8, the expected value of the ACF_1 is .8, that of the ACF_2 is .64, that of the ACF_3 is

.512, and so on. The PACF for an AR(1) model is as depicted in Panel B of Figure 9.4. The PACF has a single nonzero expected value at time lag 1 and zero expected values at all other time lags. For example, for a φ_1 value of .8, the expected value of the $PACF_1$ is .8, that of the $PACF_2$ is 0, that of the $PACF_3$ is 0, and so on. An MA(1) model has the reverse pattern: a single nonzero ACF expected value but all nonzero PACF expected values. That is, an MA(1) model has an ACF with the expected pattern in Panel B of Figure 9.4, while the PACF has the expected pattern in Panel A of Figure 9.4. The ACF and PACF expected values for more complex ARMA models combine the ACF and PACF patterns for the corresponding AR and MA models.

The second step in fitting an AR, MA, or ARMA model is the estimation phase. A regression model (e.g., Equation 9.1 or 9.2) is fit to the data, along with the AR, MA, or ARMA model identified in the first phase of the model-fitting process. Such composite models (i.e., regression model along with AR, MA, or ARMA specifications) can be fit using procedures such as maximum likelihood or generalized least squares (Maggin et al., 2011).

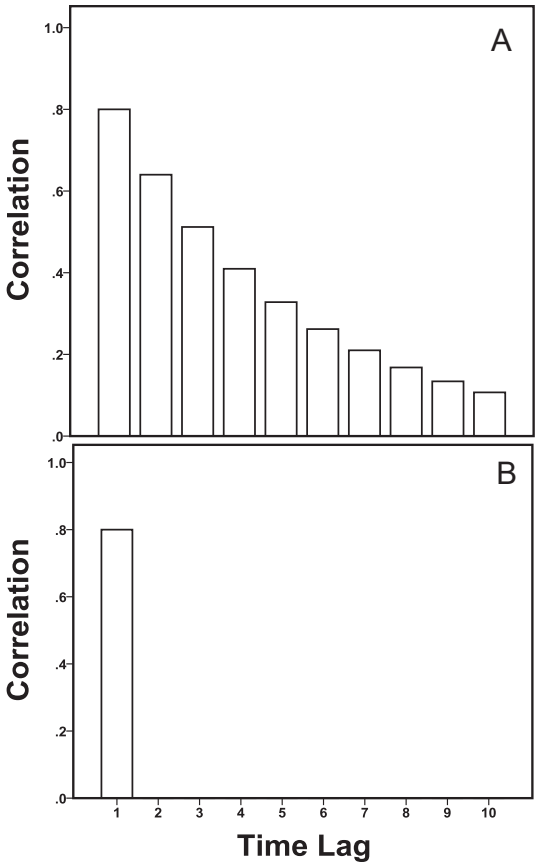


FIGURE 9.4. Panel A: An idealized autocorrelation function for an AR(1) model. Panel B: An idealized partial autocorrelation function for an AR(1) model.

The third phase in the model-fitting process is the diagnosis phase wherein the residuals from the model fit in the second phase are examined. If the correct model has been fit to the data, the residuals should be white noise, which means that the residuals are not autocorrelated, so the values of ACF and PACF all have expected values of zero. If this is not the case, the three phases are repeated. That is, a new ARMA model is chosen, fit, and diagnosed until an adequate model is found with residuals that are white noise. Estimates of treatment effects are then derived from the fit of the final model in the second phase.

Other approaches to dealing with autocorrelation are also possible. For example, Velicer and McDonald (1984) propose a method that does not require identifying the ARMA structure of the data.

9.4.2 The Number of Time Points (*J*) Is Small

When $N = 1$ and J is relatively small, Equations 9.1 and 9.2 can still be fit to the data, as can generalized additive models (Shadish et al., 2014; Sullivan et al., 2015). But the modeling of autocorrelation is more limited. It is often said that researchers need a large number of time points to conduct the three-phase model-fitting process for coping with autocorrelation that is described above. Depending on the quality of the data, as many as 50 to 100 or more time points are often said to be needed (Shadish et al., 2002; McCleary et al., 2017). With less data, the first step in the three-phase model-fitting process described above is likely to produce highly uncertain results because the ACF and PACF cannot be well estimated and it will be difficult to choose among the different models (Velicer & Harrop, 1983). Perhaps the best that can be done is to calculate a Durbin–Watson statistic to test for the presence of a lag-one autocorrelation (Huitema, 2011; Maggin et al., 2011). In any case, researchers will often forgo the three-phase process and just fit multiple models, assuming different AR, MA, and ARMA structures (including fitting an OLS regression model, which assumes there is no autocorrelation among the residuals) to see how the results compare. Researchers can feel confident in the treatment effect estimates if the results from different models converge on the same estimates for the treatment effects. If the researcher uses only a single model, the generally accepted default option is an AR(1) model.

9.5 THE STATISTICAL ANALYSIS OF DATA WHEN *N* IS LARGE

The previous sections dealt with the analysis of a single ($N = 1$) time series of data. The present section assumes that a researcher has a separate time series from each of N units, where N is large and where the treatment has been implemented with each unit. In the present section, to say N is large is to imply the design consists of data from perhaps 30 or more units, which is often required for adequate power to detect treatment

effects using the procedures to be reviewed here. The number of observational time points (J) could be relatively small: perhaps as few as eight time periods. But J could also be much larger. Such data are called **panel data.**

A widely used approach to such data are models variously called hierarchical linear models (HLMs), multilevel models, and random coefficient models (Bloom, 2003; Graham et al., 2009; Rindskopf & Ferron, 2014; Shadish et al., 2013; Singer & Willett, 2003; Somers et al., 2013). These same models were introduced in Section 4.9 on cluster-randomized experiments.

For present purposes, I consider multilevel models with just two levels. At the lower level (Level 1), a model is fit to each unit's time series of observations. A model that would allow for linear trends with a change in both level and slope for each unit would be the following:

Level 1

$$Y_{ij} = \alpha_i + (\beta_{1i}\,\text{POSTINTERVENTION}_{ij}) + (\beta_{2i}\,\text{TIME}^*_{ij}) + [\beta_{3i}\,(\text{POSTINTERVENTION}_{ij} \times \text{TIME}^*_{ij})] + \varepsilon_{ij} \tag{9.7}$$

This model is essentially the same as the model in Equation 9.1 except that there are now two subscripts on the variables to indicate that there are multiple units as well as multiple observations over time for each unit. The i subscript represents the units (so i varies from 1 to N). The j subscript represents times (so j varies from 1 to J). The Level 1 model in Equation 9.7 is fit to each unit, where

Y_{ij} is the outcome score for the ith unit at the jth time point;
$\text{POSTINTERVENTION}_{ij}$ is an indicator variable that equals 0 before the intervention and 1 afterward for the ith unit;
TIME_{ij} is a variable representing time for the ith unit: Coded 1 for the first time point, 2 for the second time point, and so on (assuming the observations are equally spaced)
TIME^*_{ij} equals (TIME_{ij} minus TIMEI_i) where TIMEI_i is the time of the intervention for the ith unit;
$\text{POSTINTERVENTION}_{ij} \times \text{TIME}^*_{ij}$ is the product of $\text{POSTINTERVENTION}_{ij}$ and TIME^*_{ij};
ε_{ij} is the residual (or how much each individual observation deviates from the ith unit's trend lines in the pretest and posttest data);
α_i is the level of the pretreatment time series at TIMEI for the ith unit;
β_{1i} is the change in level of the data from before to after the intervention at TIMEI for the ith unit (based on a comparison of the pretreatment and posttreatment trends);
β_{2i} is the slope of the observations before intervention for the ith unit; and
β_{3i} is the change in slopes from before to after the intervention for the ith unit.

The values of β_{1i} and β_{3i} are the treatment effects for each unit. The value of β_{1i} is the change in level in the *i*th unit due to the treatment, and the value of β_{3i} is the change in slope for the *i*th unit due to the treatment. If the trends in the data were curvilinear, the researcher could add polynomial terms to the model (Shadish et al., 2013). The semiparametric approach of generalized additive models is also available for fitting curvilinear trends in ITS data (Shadish et al., 2014).

While β_{1i} and β_{3i} are the treatment effects for the individual units, desired are the estimates of these treatment effects averaged across units. Such average effects are estimated at the second level of the multilevel model. The Level 2 model is

Level 2

$$\begin{aligned} \alpha_i &= \zeta_0 + r_{0i} \\ \beta_{1i} &= \zeta_1 + r_{1i} \\ \beta_{2i} &= \zeta_2 + r_{2i} \\ \beta_{3i} &= \zeta_3 + r_{3i} \end{aligned} \tag{9.8}$$

where

ζ_0 is the mean of the α_i parameters across units;
ζ_1 is the mean of the β_{1i} parameters across units;
ζ_2 is the mean of the β_{2i} parameters across units;
ζ_3 is the mean of the β_{3i} parameters across units; and
r terms are residuals (how much the parameter for each individual unit varies around the overall parameter means across units).

The change in level after the treatment is introduced averaged across units is ζ_1, and the change in slope after the treatment is introduced averaged across units is ζ_3. In other words, ζ_1 and ζ_3 are the effects of the treatments on level and slope, respectively, averaged across the units. Specifications of even more complex models can be found in Shadish et al. (2013).

A researcher can add covariates at the second level to assess how the treatment effect varies across different subclasses of units (Graham, Singer, & Willett, 2009; Singer & Willett, 2003). For example, a researcher could add a variable denoting sex to see how the changes in levels or slopes differ for males and females. In this case, the Level 1 model would remain the same and the Level 2 model would be augmented:

Level 2

$$\begin{aligned} \alpha_i &= \zeta_{00} + \zeta_{01}\,\mathrm{SEX}_i + r_{0i} \\ \beta_{1i} &= \zeta_{10} + \zeta_{11}\,\mathrm{SEX}_i + r_{1i} \\ \beta_{2i} &= \zeta_{20} + \zeta_{21}\,\mathrm{SEX}_i + r_{2i} \\ \beta_{3i} &= \zeta_{30} + \zeta_{31}\,\mathrm{SEX}_i + r_{3i} \end{aligned} \tag{9.9}$$

where

SEX_i is an indicator variable coded 0 for males and 1 for females;
ζ_{00} is the mean of the α_i parameters for males;
ζ_{01} is the difference in the means of the α_i parameters between males and females;
ζ_{10} is the mean of the β_{1i} parameters for males;
ζ_{11} is the difference in the means of the β_{1i} parameters between males and females;
ζ_{20} is the mean of the β_{2i} parameters for males;
ζ_{21} is the difference in the means of the β_{2i} parameters between males and females;
ζ_{30} is the mean of the β_{3i} parameters for males;
ζ_{31} is the difference in the means of the β_{3i} parameters between males and females; and
r terms are residuals (how much each parameter varies around the group means).

The value of ζ_{10} is the estimate of the average change in level for males due to the treatment. The value of ζ_{11} reveals how much males and females differ on average in the change in level due to the treatment. The value of ζ_{30} is the estimate of the average change in slope for males due to the treatment. The value of ζ_{31} reveals how much males and females differ on average in the change in slope due to the treatment.

Researchers can take account of autocorrelations among the observations at the level of the units (Level 1) if the sample size is sufficient (Shadish et al., 2013). An AR(1) model is usually the default option when autocorrelation is present, but a more complex model cannot be specified. There is no need to take account of autocorrelation at the second level because there are no time-based variables at the second level.

Multilevel models are very flexible. For example, because all the variables in Equation 9.7 have two subscripts, the values of $POSTINTERVENTION_{ij}$ and $TIME^*_{ij}$ can vary across individuals. This means each unit could have an intervention that begins at a different point in time, each unit could have a different number of observations over time, and the observations for different units could be spaced at different time intervals. Multilevel models automatically accommodate missing data, but that does not mean missing data will not introduce biases. Multilevel models assume that missing data are MAR or MCAR (see Section 4.8).

A researcher can also perform multilevel analyses of time-series data using growth curve structural equation models (Duncan & Duncan, 2004a, 2004b; Schoemann, Rhemtulla, & Little, 2014). With the growth curve approach, researchers can have the treatment effect variables predict and be predicted by other variables. A growth curve model can also include latent variables and multiple growth curves for multiple outcome variables in the same model. The growth curve approach also provides tests of goodness of fit that are not available with the multilevel approach. However, the growth curve approach is not as flexible as the multilevel approach when data are available at different times and the treatment is introduced at different times for different units. Other parametric approaches to statistical analysis, besides multilevel models and growth curve models, are also possible (Algina & Olejnik, 1982; Algina & Swaminathan, 1979).

9.6 THREATS TO INTERNAL VALIDITY

As mentioned in Section 9.1, the ITS design is an extension of the pretest–posttest design. Because of the way the ITS design extends the pretest–posttest design, the ITS design is far less susceptible to most of the threats to internal validity that can plague the pretest–posttest design. Chapter 6 introduced eight threats to internal validity that provide alternative explanations for the results of a pretest–posttest design. In practice, most of these threats can be ruled out in the basic ITS design or with the supplements to the basic ITS design, as described in Sections 9.7, 9.8, and 9.9. That the ITS design is likely susceptible to far fewer threats to validity than the pretest–posttest design is the reason the ITS design generally produces more credible results than does the pretest–posttest design.

9.6.1 Maturation

Maturation can produce a trend in a time series of observations. The threat to internal validity of maturation is controlled in an ITS design by using the multiple observations before the treatment is introduced. With multiple observations, the maturational trend in the pretreatment observations can be modeled and projected forward in time to be compared with the trend in the posttreatment observations. In this way, the effects of maturation are modeled and taken into account because the treatment effect is assessed against the background trend in the data that is due to maturation.

Of course, it is possible the maturational trend in the data will be modeled incorrectly. Incorrectly modeling the maturational trend can lead to bias in the estimates of the treatment effect. For example, fitting a linear trend to data that have curvilinear maturation can produce biased treatment effect estimates (just as fitting a linear trend to curvilinear data in a regression discontinuity design can produce biased treatment effect estimates—see Section 8.3). But proper modeling allows the threat of maturation to be taken into account. So the ITS design provides a mechanism for controlling the threat of maturation that is not available in the simple pretest–posttest design.

9.6.2 Cyclical Changes (Including Seasonality)

Seasonality and cyclical changes can arise because, for example, months share similarities across years. Seasonality is not restricted to yearly cycles, however. Cycles can also arise across days within the week (Liu & West, 2015) and across hours within the day. Cyclical changes are like maturation in that ITS models provide a means to take account of cyclical patterns (including with ARMA models). If the pretreatment time series is sufficiently long, cyclical changes can be modeled and their effects removed from the observations before the treatment effect is estimated.

9.6.3 Regression toward the Mean

If regression toward the mean is to arise, the observations preceding the implementation of the treatment (at the time when participants are selected to receive the treatment) should be either higher or lower than average (when assessed based on the entire pretreatment series of observations). By including many more pretreatment observations than just a few, the researcher can determine if regression toward the mean is plausible (Campbell & Ross, 1968). If the observations preceding the treatment implementation (at the time when participants are selected to receive the treatment) are not unusual compared to even earlier or later pretreatment observations, regression toward the mean is usually an implausible explanation for the results. Researchers should also assess the plausibility of regression toward the mean by investigating the selection mechanism to see if units have been selected, either directly or indirectly, so they have higher or lower observations prior to implementation of treatment.

9.6.4 Testing

The effects of testing generally wear off after a few observations. Because the ITS design has multiple observations collected before the treatment is introduced, the design is generally not susceptible to biases due to testing effects. Even if the effects of testing are still present after multiple observations, their pattern of effects can often be assessed, much like maturation, and taken into account by modeling the trend in the pretreatment observations.

9.6.5 History

History effects are due to external events that occur at the same time the treatment is introduced. Hence, history effects threaten the internal validity of a basic ITS design because an interruption in the time series could be due either to the treatment or to history. The shorter the time lag between observations and the shorter the time it takes for the treatment to have its effects after it is introduced, the less time there is for history effects to occur and therefore to masquerade as treatment effects. Conversely, if the delay between treatment introduction and the onset of a treatment effect is unknown and potentially long, the more opportunity there is for a history effect to arise and bias the estimate of a treatment effect. Researchers should consider adding supplements to the basic ITS design to help rule out threats to internal validity due to history, as explained in Sections 9.7, 9.8, and 9.9.

9.6.6 Instrumentation

A threat to internal validity due to instrumentation arises when the measuring instrument changes when the treatment is introduced. The basic ITS design is susceptible to

threats to internal validity due to instrumentation, but adding multiple interventions or a comparison time series can help rule out the threat, as explained in Sections 9.7, 9.8, and 9.9. For example, finding no evidence of an effect due to instrumentation in an adequate comparison time series supports the conclusion that instrumentation effects are not present in the experimental time series.

The threats of history and instrumentation can also be assessed by collecting auxiliary measurements, for example, by examining newspapers or other records and by talking with knowledgeable witnesses at the local scene. After diligently searching for evidence of history effects or changes in instrumentation, lack of evidence would suggest lack of bias due to these threats to internal validity.

9.6.7 Selection Differences (Including Attrition)

Changes in the composition of the units (i.e., selection differences) at the point of the intervention can be a threat to the internal validity of a basic ITS design. For example, if a hospital introduces a new medical service at the same time the experimental intervention is introduced, numerous outcome variables could be affected such as physician visits, hospital stays, and assessments of comorbidity because of differences in the number and types of patients who seek treatment from the hospital. Alternatively, some participants may drop out of a study because the treatment is too demanding of their time or is not to their liking for other reasons. Again, adding supplements such as a comparison time series (perhaps a comparison time series using a nonequivalent dependent variable) that is susceptible to the same threats to internal validity can help rule out the threat, as explained in Sections 9.7, 9.8, and 9.9. When some units drop out of the study, another solution is to conduct the analysis using data only from those who did not drop out to make the composition of the sample the same after the treatment was introduced as it was before the treatment was introduced. If that is not possible, look for changes, from before the treatment to after the treatment, in the background characteristics of the participants. The fewer and smaller such differences are, the less plausible becomes a threat to validity due to composition changes.

9.6.8 Chance

Chance differences can introduce a change in the time series of observations coincident with the introduction of the treatment. Statistical models treat unexplained variability in outcomes as if they were due to random or chance influences. Such random variation is modeled using statistical procedures as described in Sections 9.4 and 9.5. As a result, a change in the time series of observations due to chance is ruled out using statistical procedures. Nonetheless, the presence of random variability reduces the power and precision of the statistical analyses (as is true for all designs).

9.7 DESIGN SUPPLEMENTS I: MULTIPLE INTERVENTIONS

The basic ITS design can be supplemented in myriad ways. The primary purpose of supplements is to help control threats to internal validity. Indeed, some methodologists argue that, unless the treatment effect happens abruptly and is large relative to variation in the observations over time, results from the basic ITS design are unreliable without supplements. In the present section, I consider three different types of supplements involving adding multiple treatment interventions. The analyses of data from these supplements are themselves supplements of the analyses in Sections 9.4 and 9.5 (also see Huitema, 2011).

9.7.1 Removed or Reversed Treatment Designs

One way to supplement the basic ITS design is to remove or reverse the treatment after it has been introduced. The design (called a reversal design) is diagrammed thusly:

$$O_1\ O_2\ O_3\ O_4\ O_5\ X\ O_6\ O_7\ O_8\ O_9\ O_{10}\ \cancel{X}\ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}$$

where $\cancel{X}$ denotes the removal or reversal of the treatment. In words, the treatment is introduced, allowed to operate for a period of time, and then removed or reversed, with observations taken both before and after each of these two interventions. The expectation in this design is for an interruption in the time series when the treatment is first introduced and then an interruption in the opposite direction when the treatment is removed or reversed. For example, the U.S. General Accounting Office (1991) used the design to assess the effects of a federal law mandating that motorcycle riders wear helmets. The law was introduced in 1965 and then repealed in 1975. Fatalities due to motorcycle accidents decreased following the introduction of the law and increased following its repeal.

This design clearly requires a treatment whose effects can be removed or reversed once the treatment has been introduced. For example, the effects of behavioral reinforcements might continue as long as the reinforcements continue but then they stop after the reinforcements are removed. But not all treatments are of this nature. A successful intervention to teach a participant to ride a bicycle cannot be removed or reversed. Once taught to ride a bicycle, people do not usually unlearn those skills just because instruction ends.

The added complexity of this supplemented design, compared to a basic ITS design, helps rule out threats to internal validity such as history and instrumentation. If the treatment were effective, each of the two treatment interventions in the reversal design would produce an interruption in the time series. Threats to internal validity, such as history and instrumentation, could explain a pattern of repeated interruptions only if there were two threats operating; one when the treatment was introduced and another when the treatment was removed or reversed. In addition, the threats

to internal validity would have to have effects in the opposite directions because the removal or reversal of the treatment would have an effect in the opposite direction of the initial treatment effect. To the extent the pattern of data is as predicted by a treatment effect, threats to internal validity are rendered less plausible, than in the basic ITS design, as explanations for the results. The relative advantage of the reversal design can be explained in the following manner. Because it is less likely that more than one threat to internal validity would be present and account for the opposing results in the reversal design than that a single threat to internal validity would be present to account for the results in the basic ITS design, threats to internal validity are rendered less plausible in the more complex design than in the simpler one. See Chapter 12 for further discussion of the benefits of adding features, such as removed or reversed treatments, that lead to complex patterns of results.

9.7.2 Repeated Treatment Designs

A design can have both a removed treatment and, subsequently, a repeated treatment. Such a design is often called just a repeated treatment design and would look like the following:

$$O_1\ O_2\ O_3\ O_4\ O_5\ X\ O_6\ O_7\ O_8\ O_9\ O_{10}\ \cancel{X}\ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}\ X\ O_{16}\ O_{17}\ O_{18}\ O_{19}\ O_{20}$$

In words, the treatment is introduced, then removed, and then reintroduced—with observations collected before and after each of the three interventions. Of course, if the effect of the treatment is transient, the treatment could simply be repeated without having to be explicitly removed. Such a design would be diagrammed as

$$O_1\ \ O_2\ \ O_3\ \ O_4\ \ O_5\ \ X\ \ O_6\ \ O_7\ \ O_8\ \ O_9\ \ O_{10}\ \ X\ \ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}$$

An example would be a medication where the effects wear off but the underlying condition that the medication treats remains over time (such as pain after surgery).

In either form of the design, there would be multiple changes or transitions due to the treatment. The first implementation of the treatment would introduce an effect compared to the pretreatment baseline of observations. The treatment effect would subsequently disappear (either because the effect is transient or because the treatment is removed). Then the treatment effect would reappear with the reintroduction of the treatment (perhaps to wear off again if the treatment has a transient effect).

When the assumptions of the repeated treatment design are met, the design can help rule out threats to internal validity such as history and instrumentation, in the same way that such threats are ruled out in the reversal treatment design. That is, with the repeated treatment design, if history events or instrumentation effects are to account for the pattern of outcomes, they would have to occur more than once, happening with the treatment each time the treatment effect is altered. That is, they would

occur at the time the treatment is initially introduced, at the time the treatment is removed or wears off, and then again at the time the treatment is reintroduced (and perhaps also after the treatment wears off a second time). The threats to internal validity would again have to have effects in opposite directions (just as the treatment introduction and removal are presumed to produce effects in opposite directions). That is, a history or instrumentation threat to internal validity would have to operate multiple times in specific directions to account for the multiple transitions in the repeated treatment design. Because it is less likely that multiple biases would occur at the required times and in required directions than that a single bias would be present in a given direction, threats to internal validity are rendered less plausible in the more complex design than in the basic ITS design (assuming the predicted complex pattern of treatment effects is in evidence). Note, too, that the repeated treatment design can increase credibility even when there is an unknown delay in the onset of a treatment effect. With the single-treatment implementation, an unknown delay in the onset of a treatment effect allows more time for threats to internal validity to arise and thereby account for the results. But an interruption in the time series that occurs at the same delayed time lag after multiple treatment implementations is generally more plausibly explained as being due to the treatment than to repeated threats to validity.

Note that a repeated treatment design could be elaborated even further by having more instances of repeated and removed treatments. The more times the treatment is repeated and removed, the more credibly the results can be attributed to the treatment rather than to threats to internal validity.

9.7.3 Designs with Different Treatments

In the repeated treatment design, the same treatment is implemented at different times. Alternatively, a different treatment or the same treatment at different doses could be introduced for purposes of comparison. For example, a design could take the following form:

$$O_1\ O_2\ O_3\ O_4\ O_5\ X_A\ O_6\ O_7\ O_8\ O_9\ O_{10}\ \cancel{X}_A\ O_{11}\ O_{12}\ O_{13}\ O_{14}\ O_{15}\ X_B\ O_{16}\ O_{17}\ O_{18}\ O_{19}\ O_{20}$$

where X_A is the initial treatment or the initial dose of a treatment, $\cancel{X}_A$ is the removal of that treatment or dose, and X_B is a different treatment or a different dose of the same treatment. There are more possibilities as well, all with much the same advantages as with previously described supplements.

9.8 DESIGN SUPPLEMENTS II: BASIC COMPARATIVE ITS DESIGNS

Another supplement to the basic ITS design is the addition of a comparison time series of observations, where the comparison time series does not receive the treatment given in the treatment condition. For example, the comparison time series could consist of a

nonequivalent group of participants who do not receive the treatment. Such a design would be diagrammed thusly:

NR: O_1 O_2 O_3 O_4 O_5 O_6 X O_7 O_8 O_9 O_{10} O_{11} O_{12}

NR: O_1 O_2 O_3 O_4 O_5 O_6 O_7 O_8 O_9 O_{10} O_{11} O_{12}

The first row in the diagram represents the experimental time series where a participant or group of participants receives the experimental treatment. The second row in the diagram represents the comparison time series where a different participant or group of participants receives a comparison treatment (such as a standard treatment or no treatment at all). The "NR:" in the diagram indicates that participants are assigned to these two treatment conditions nonrandomly. Of course, the participants could be assigned to treatment conditions at random, but then the design would be a randomized experiment. (Bloom, 2005b, provides an example.)

Other types of comparison time series are also possible. Instead of a comparison time series derived from different participants, the comparison time series could be derived from the same participants but for different times, settings, or outcome measures (see Chapters 10 and 11). When different outcome measures are used, for example, the design is said to involve nonequivalent dependent variables. In the present section, I focus on analysis of data from a comparative interrupted time series (CITS) where the comparison time series comes from a nonequivalent group of participants rather than designs where the comparison time series comes from different times, settings, or outcome measures (see Chapter 11). The logic of the analysis of designs with different types of comparison time series is much the same, though with designs other than nonequivalent group of participants there may be dependencies across the treatment and comparison time series (see Huitema, 2011).

In some literatures, ITS designs with an added comparison time series of observations are called difference-in-differences (DID) designs (Angrist & Pischke, 2009, 2015; Wing, Simon, & Bello-Gomez, 2018). Other authors distinguish between DID designs in the context of the nonequivalent group design (see Chapter 7) and DID designs in the context of ITS designs, which they call comparative interrupted time-series (CITS) designs (St. Clair & Cook, 2015; St. Clair, Hallberg, & Cook, 2016; Somers et al., 2012). For clarity, I use the CITS label for any ITS design with an added comparison time series.

A primary purpose of adding the comparison time series in a CITS design is to rule out threats to internal validity such as those due to history and instrumentation. The comparison time series is selected to be susceptible to the same threats to internal validity as the treatment group at the time the treatment is introduced. But the comparison time series is not susceptible to the treatment effect. An estimate is obtained from the comparison time series as if the treatment had been introduced into that series at the same time as the treatment was introduced into the experimental series. To the

extent that the estimate (a pseudo treatment effect) from the comparison series is not substantially different from zero, the researcher concludes that the effects of the shared threats to internal validity in the experimental series is not substantially different from zero. Then any difference in the experimental series that is substantially different from zero is assumed to be due to the treatment rather than to the shared threats to internal validity. By this logic, a comparison time series of observations could be used to rule out different threats to internal validity all at once (such as threats due to history and instrumentation) to the extent that the comparison time series is susceptible to these different threats to internal validity.

Guerin and MacKinnnon (1985) provide an example of a CITS design with a nonequivalent group comparison time series. The study assessed the effects on fatalities of a state law requiring that infants from 0 to 3 years of age wear passenger restraints in cars. The experimental time series measured traffic fatalities of children in this age group. The comparison time series measured traffic fatalities of children aged 4–7 years. Because the law did not require passenger restraints for the older children, the treatment was presumed to have little, if any, effect in that age group. But most plausible threats to internal validity such as due to history and instrumentation should affect the two time series similarly. That the experimental time series showed a drop in fatalities when the law was introduced, while the comparison time series showed no interruption, is evidence that the treatment was effective rather than that the results were due to a shared threat to internal validity.

McKillip (1992) provides an example of a CITS design with a nonequivalent dependent variable where the effects of an educational media campaign to increase the responsible use of alcohol on a college campus was assessed. Survey questions evidenced increased awareness of responsible use of alcohol following the campaign. Questions about nutrition and stress reduction (i.e., nonequivalent dependent variables that should have exhibited no treatment effect but would be susceptible to the same history effects) showed no comparable changes in awareness, thereby ruling out history effects as plausible alternative explanations for the results.

Adding a comparison time series to a CITS design has the further advantage that a comparison series can increase the credibility with which a change in slope in the experimental series is attributed to the effect of the treatment rather than to a misspecified curvilinear trend in the data. Such increased credibility is obtained when the comparison series has the same pretreatment trend as in the experimental series and when the comparison series is hypothesized to have the same trend in the posttreatment data as in the experimental series in the absence of a treatment effect. Then any difference (including changes in slope) between the experimental and comparison time series in posttreatment trends can be attributed to the effect of the treatment. Similarly, adding a comparison time series to a CITS design can help the researcher assess treatment effects that appear gradually over time. As noted before, the gradual onset of a treatment effect may be difficult to distinguish from a curvilinear pattern of change that might arise when there is no treatment effect. To the extent that the trends in the comparison

time series of observations closely parallels the trends in the experimental time series of observations before the treatment is introduced, deviations between the two series after the treatment is introduced, including gradual changes in the experimental series, can be more credibly attributed to a treatment effect rather than, say, to a curvilinear pattern of maturation. Yet one more advantage of the CITS design is that the results can be more powerful and precise than the results from a basic ITS design.

Of course, the CITS design is not perfect. For example, it can still suffer from threats to internal validity, such as those due to differential history or differential instrumentation where the threat to internal validity has an effect only in the experimental time series. Threats such as these are judged implausible to the extent that the treatment and comparison series are similar except for differences in the intended interventions.

The statistical analysis of data from a CITS is a generalization of the analysis of data from the basic ITS. As always, the analysis begins by plotting the data. Because statistical models for CITS designs can get very complex when fitting data from both treatment and comparison conditions simultaneously, I would begin statistical analysis by fitting models to the time-series data from each treatment condition separately, using models such as those in Sections 9.4 and 9.5. Then only after I had a feel for the results in each treatment condition separately would I combine the data from both treatment conditions into a single analysis, as outlined below. Analyses can take account of autocorrelations among the residuals (see Section 9.4) if the time series are sufficiently long (e.g., Shadish et al., 2013). As in Sections 9.4 and 9.5, I will discuss statistical analyses in two sections: when $N = 1$ in each treatment condition and when N is relatively large in each treatment condition.

9.8.1 When $N = 1$ in Each Treatment Condition

Because, as I said, the simultaneous analysis of data from both the treatment and comparison time series can be complex, I will start with simpler simultaneous analyses and progress to more complex simultaneous analyses. The first two models do not allow for pseudo treatment effects in the comparison time series. That is, the first two models do not permit changes in either level or slope in the comparison time series that are due to threats to internal validity that occur when the treatment is introduced in the treatment time series. These models are introduced as a way to build up gradually to the very complex third model in which pseudo treatment effects (due to threats to internal validity) are allowed.

9.8.1.1 CITS Model 1

Suppose separate analyses of the data from the treatment and comparison conditions revealed the following three results. First, the treatment produced a change in level but not in slope, so the trend in the data in the treatment condition was the same before and after the intervention. Second, the slope in the comparison condition was the same as in

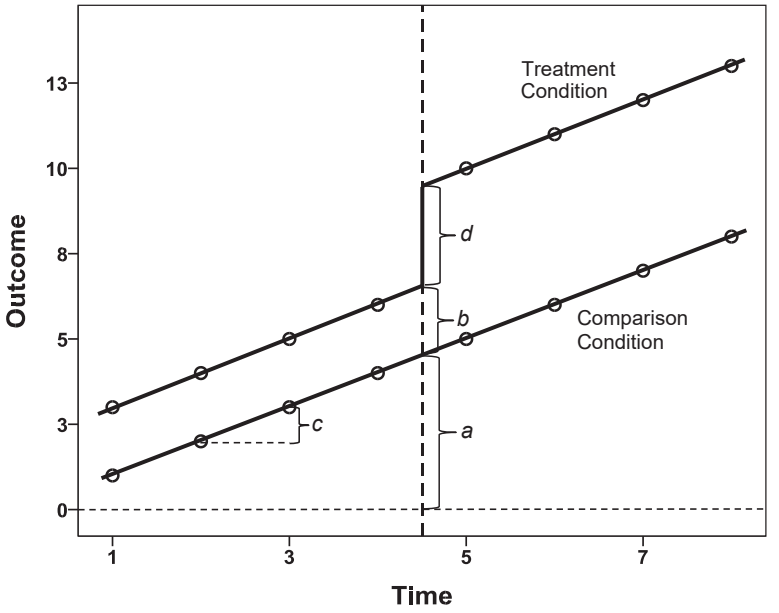


FIGURE 9.5. Idealized results for a comparative interrupted time-series design that fits Equation 9.10 in the text.

the treatment condition and was the same before and after the intervention. Third, there was no pseudo treatment effect in the comparison time series. Such a pattern of results is depicted in Figure 9.5. In this figure, time is plotted on the horizontal axis, and outcomes are plotted on the vertical axis. The lines plotted in the figure are idealized, fitted regression lines; scatter around these lines is omitted. The observed data (were they to be plotted but, for simplicity, they are not) would vary around these idealized regression lines. The dots on these regression lines indicate when observations were taken. For ease of presentation, observations are drawn as if taken at just eight time points, but there could be many more observations. The vertical dashed line indicates when the treatment intervention was introduced. In this case, the treatment was introduced halfway between the times of the fourth and fifth observations.

To model the results depicted in Figure 9.5, the following equation could be fit to the data:

$$Y_{ij} = \alpha + (\beta_1 \text{ CONDITION}_{ij}) + (\beta_2 \text{ TIME}^*_{ij}) + [\beta_3 \text{ CONDITION}_{ij} \times \text{POSTINTERVENTION}_{ij}] + \varepsilon_{ij} \quad (9.10)$$

where

i is a subscript representing the treatment condition;
j is a subscript representing time (where *j* varies from 1 to *J*);

Y_{ij} is the outcome score for the ith treatment condition at the jth time point;
CONDITION_{ij} is an indicator variable that equals 1 if the observation is from the treatment condition and 0 if the observation is from the comparison condition;
TIME_{ij} is a variable representing time: coded 1 for the first observation, 2 for the second observation, and so on (assuming the observations are equally spaced);
TIME^*_{ij} equals (TIME_{ij} minus TIMEI) where TIMEI is the time of the intervention;
$\text{POSTINTERVENTION}_{ij}$ is an indicator variable that equals 1 if the observation takes place after TIMEI and is 0 otherwise;
$\text{CONDITION}_{ij} \times \text{POSTINTERVENTION}_{ij}$ is the product of CONDITION_{ij} and $\text{POSTINTERVENTION}_{ij}$;
ε_{ij} is the residual that reveals how much observations deviate from the trend lines;
α is the level of the observations in the comparison condition at TIMEI (labeled *a* in Figure 9.5);
β_1 is the difference in levels between the treatment and comparison conditions in the preintervention data at TIMEI (labeled *b* in Figure 9.5);
β_2 is the common slope of the trends in the observations in the comparison and treatment conditions both before and after the intervention (labeled *c* in Figure 9.5); and
β_3 is the change in the level in the treatment condition from before to after the intervention at TIMEI (labeled *d* in Figure 9.5).

The estimate of β_3 is the estimate of the treatment effect, which is a change in level. The analysis using Equation 9.10 could be called a difference-in-differences (DID) analysis because the effect of the treatment could be estimated as a difference in differences. That is, the effect of a change in level (β_3) could be estimated as the difference between (1) the average pretreatment-to-posttreatment difference in the comparison condition and (2) the average pretreatment-to-posttreatment difference in the treatment condition. Note that this DID analysis assumes a common trend over time in the treatment and comparison time series, both before and after the treatment is introduced. This will not always be a reasonable assumption and should be checked in the data. If the trends are not the same, Equation 9.10 likely gives biased estimates of the treatment effect.

9.8.1.2 CITS Model 2

Suppose the treatment causes a change in both the level and slope in an experimental time series, but the pretreatment trends in the data are the same in the experimental and comparison time series and there is no pseudo treatment effect in the comparison time series due to threats to internal validity. Such a pattern is depicted in Figure 9.6 using the same conventions as in Figure 9.5.

The following model would fit the idealized results in Figure 9.6:

$$Y_{ij} = \alpha + (\beta_1 \text{ CONDITION}_{ij}) + (\beta_2 \text{ TIME}^*_{ij}) + (\beta_3 \text{ CONDITION}_{ij} \times \text{POSTINTERVENTION}_{ij}) + (\beta_4 \text{ CONDITION}_{ij} \times \text{TIME}^*_{ij} \times \text{POSTINTERVENTION}_{ij}) + \varepsilon_{ij} \quad (9.11)$$

where

α is the level of the observations in the comparison condition at TIMEI (labeled *a* in Figure 9.6);

β_1 is the difference in levels between the treatment and comparison conditions in the preintervention data at TIMEI (labeled *b* in Figure 9.6);

β_2 is the slope of the trends in the observations in (1) the comparison condition both before and after the intervention and (2) the treatment condition before the intervention (labeled *c* in Figure 9.6);

β_3 is the change in level in the treatment condition from before to after the intervention at TIMEI (labeled *d* in Figure 9.6);

β_4 is the change in slope in the treatment condition from before to after the intervention (*e* in Figure 9.6 = $\beta_2 + \beta_4$); and

The variables in the model are defined as in Equation 9.10 (or are products of variables defined in Equation 9.10).

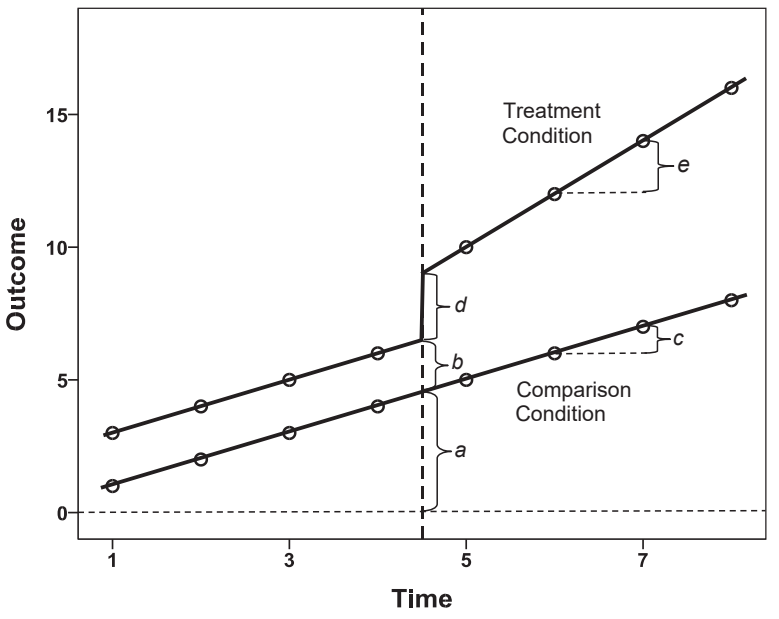


FIGURE 9.6. Idealized results for a comparative interrupted time-series design that fits Equation 9.11 in the text.

The treatment effect estimates are β_3 and β_4. The estimate of β_3 reveals how much the treatment changes the level of the experimental time series when the treatment is introduced at TIMEI. The estimate of β_4 reveals how much the treatment changes the slope of the regression lines in the treatment condition from before to after the treatment is introduced. For example, if the slope of the regression line in the treatment condition before the intervention is 1 and is 1.5 afterward, β_4 is equal to the difference between 1 and 1.5, which is .5.

9.8.1.3 CITS Model 3

The data could take an even more complex pattern than in Figure 9.6. If trends are all linear, a more complex pattern is given in Figure 9.7. In Figure 9.7, there could be changes in both level and slope in both the treatment and comparison time series. That means there are pseudo treatment effects in the comparison time series due to threats to internal validity. These pseudo effects are allowed to cause a change in both the level and slope in the comparison time series. The treatment effects are estimated as changes in level and slope in the treatment time series that are larger than the changes in level and slope in the comparison time series.

For idealized data as in Figure 9.7, the following model could be fit (Linden, 2015; Wong, Cook, & Steiner, 2015):

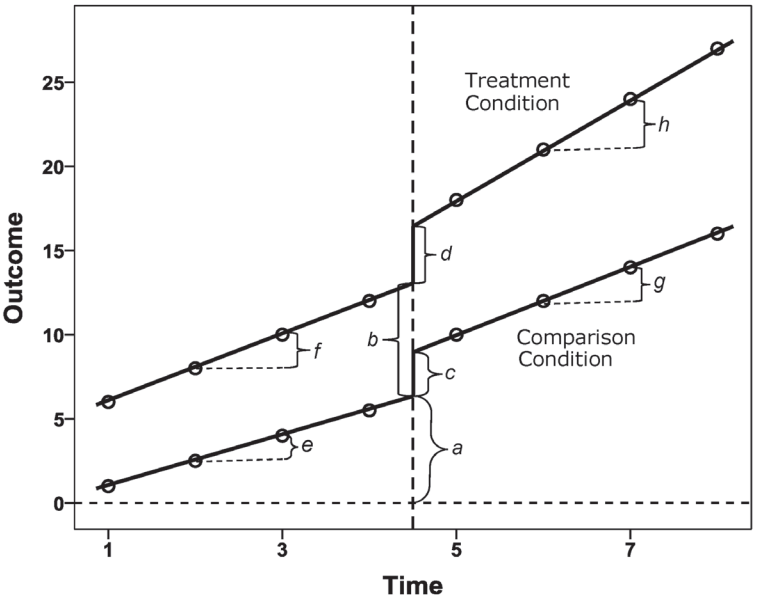


FIGURE 9.7. Idealized results for a comparative interrupted time-series design that fits Equation 9.12 in the text.

$$
\begin{aligned}
Y_{ij} = {} & \alpha + (\beta_1\ \text{CONDITION}_{ij}) + (\beta_2\ \text{POSTINTERVENTION}_{ij}) \\
& + (\beta_3\ \text{CONDITION}_{ij} \times \text{POSTINTERVENTION}_{ij}) \\
& + (\beta_4\ \text{TIME}^*_{ij}) + (\beta_5\ \text{CONDITION}_{ij} \times \text{TIME}^*_{ij}) \\
& + (\beta_6\ \text{POSTINTERVENTION}_{ij} \times \text{TIME}^*_{ij}) \\
& + (\beta_7\ \text{CONDITION}_{ij} \times \text{POSTINTERVENTION}_{ij} \times \text{TIME}^*_{ij}) + \varepsilon_{ij}
\end{aligned}
\tag{9.12}
$$

where

α is the level of the observations in the pretreatment data in the comparison condition at TIMEI (labeled *a* in Figure 9.7);

β_1 is difference in pretreatment level between the treatment and comparison conditions at TIMEI (labeled *b* in Figure 9.7);

β_2 is the change in level in the comparison condition at TIMEI (labeled *c* in Figure 9.7);

β_3 is the change in level in the treatment condition from before to after the intervention at TIMEI after taking account of β_2 (*d* in Figure 9.7 is $\beta_2 + \beta_3$);

β_4 is the slope in the pretreatment data in the comparison condition (labeled *e* in Figure 9.7);

β_5 is the difference in the pretreatment slopes between the treatment and comparison conditions (*f* in Figure 9.7 = $\beta_4 + \beta_5$);

β_6 is the change in slope from pretreatment to posttreatment in the comparison condition (*g* in Figure 9.7 = $\beta_4 + \beta_6$);

β_7 is the change in slope from pretreatment to posttreatment in the treatment condition above and beyond (1) the initial slope of pretreatment data in the comparison condition (β_4), (2) the difference in pretreatment slope between the treatment and comparison conditions (β_5), and (3) the change in slope from the pretreatment to posttreatment in the comparison condition (β_6) (*h* in Figure 9.7 = $\beta_4 + \beta_5 + \beta_6 + \beta_7$); and

The variables in the model are defined as in Equation 9.10 (or are products of variables defined in Equation 9.10).

The estimates of β_3 and β_7 are the estimates of the treatment effects. The value of β_3 is the effect of the treatment on the change in level in the treatment time series above and beyond the change in level in the comparison time series. For example, suppose the change (from pretest to posttest) in level in the comparison time series (β_2) is 1 (which is *c* in Figure 9.7) and the change in level (from pretest to posttest) in the treatment time series is 1.5 (which is *d* in Figure 9.7). Then β_3 is equal to .5 (which is 1.5 minus 1) because .5 is the change in level in the treatment time series above and beyond the change in level in the comparison time series. Similarly, β_7 is the effect of the treatment on the change in slope in the treatment time series after taking account of (1) the slope in the pretreatment data in the comparison condition (β_4), (2) the difference in

the pretreatment slopes between the treatment and comparison conditions (β_5), and (3) the change in slope from pretreatment to posttreatment in the comparison condition (β_6). In other words, if the treatment had no effect on the slope, you should predict the posttreatment slope in the treatment condition to be equal to $\beta_4 + \beta_5 + \beta_6$. For example, suppose (for the sake of a clear example) the slope in the treatment condition after the intervention (*h* in Figure 9.7) is equal to 8. Further suppose (1) the slope in the pretreatment data in the comparison condition (β_4) is equal to 3, (2) the difference in the pretreatment slopes between the treatment and comparison conditions (β_5) is equal to 2, and (3) the change in slope from pretreatment to posttreatment in the comparison condition (β_6) is equal to 1. Then β_7 is equal to 2 (which is $8 - [3 + 2 + 1]$) because 2 is how much the slope in the treatment time series changes after the intervention beyond what is expected given all the other slopes. The estimates of β_2 and β_6 in Equation 9.12 are estimates of pseudo treatment effects in the comparison time series. The effects of the treatment in the experimental time series are estimated as differences that are at least above and beyond the pseudo effects in the comparison time series.

I might note that finding a small value for the estimate of β_5 (which is the difference in pretreatment slopes between the treatment and comparison time series) could increase the researcher's confidence in drawing a credible contrast between the comparison time series and the treatment time series. The same holds for finding small values for estimates of β_2 and β_6, which are pseudo effects. Zero values for all three of these parameters would convert Equation 9.12 into Equation 9.11, which would make the interpretation of the remaining model parameters both simpler and more credible (see Section 9.8.3).

More complex patterns in the trends would require more complex models than in Equations 9.10 to 9.12. For example, polynomial terms and polynomial interactions could be added to the models. In addition, the effects of the treatments could be more complex than the abrupt sustained effects in Equations 9.10–9.12 and could be modeled with more complex equations.

9.8.2 When *N* Is Large in Each Treatment Condition

The models in Section 9.5 for use with the basic ITS when *N* is large can be generalized to apply when a comparison time series is added. I will present a two-level hierarchical linear model (HLM) approach as in Section 9.5 (Shadish et al., 2013). Other approaches are also possible; Angrist and Pischke (2009, 2015) present an analysis from the DID tradition. Both Jacob, Somers, Zhu, and Bloom (2016) and Somers et al. (2013) present an alternative HLM approach (which includes matching participants from the two treatment conditions based on pretreatment covariates). But I believe the easiest model to understand is the one I present here.

At Level 1 in the HLM model, a regression is fit to each participant's scores individually. For each individual, the model in Equation 9.13 allows for a linear slope in the

pretreatment data, a change in level from pretest to posttest, and a change in slope from pretest to posttest. The Level 1 specification is the following:

Level 1

$$Y_{ij} = \alpha_i + (\beta_{1i}\ \text{POSTINTERVENTION}_{ij}) + (\beta_{2i}\ \text{TIME}^*_{ij}) + (\beta_{3i}\ \text{POSTINTERVENTION}_{ij} \times \text{TIME}^*_{ij}) + \varepsilon_{ij} \quad (9.13)$$

where

i represents the units so i varies from 1 to $(N_1 + N_2)$ where N_1 is the sample size in the comparison condition and N_2 is the sample size in the treatment condition);
j represents time so j varies from 1 to J;
Y_{ij} is the outcome score for the ith unit at the jth time point;
$\text{POSTINTERVENTION}_{ij}$ is an indicator variable that equals 0 before the intervention and 1 afterward for the ith unit;
TIME_{ij} is a variable representing time for the ith unit: coded 1 for the first time point, 2 for the second time point, and so on (assuming the observations are equally spaced in time);
TIME^*_{ij} equals (TIME_{ij} minus TIMEI_i) where TIMEI_i is the time of the intervention for the ith unit;
$\text{POSTINTERVENTION}_{ij} \times \text{TIME}^*_{ij}$ is the product of $\text{POSTINTERVENTION}_{ij}$ and TIME^*_{ij};
ε_{ij} is the residual (or how much each individual observation deviates from the trend lines in the pretreatment and posttreatment data);
α_i is the level of pretreatment time series at TIMEI_i for the ith unit;
β_{1i} is the change in level of the data from before to after the intervention at TIMEI_i for the ith unit (based on a comparison of the pretreatment and posttreatment trends);
β_{2i} is the slope of the observations before the intervention for the ith unit; and
β_{3i} is the change in slopes from before to after the intervention for the ith unit.

Equation 9.13 is essentially the same as Equation 9.7. The values of β_{1i} and β_{3i} are the treatment effects (or pseudo treatment effects) for each unit. The value of β_{1i} is the change in level in the ith unit due to the treatment (or to a pseudo treatment effect), and the value of β_{3i} is the change in slope for the ith unit due to the treatment effect (or to a pseudo treatment effect). If the trends in the data are curvilinear, the researcher could add polynomial terms to the model.

The parameters in Equation 9.13 are derived for, and differ across, each participant. But what is desired is the average of these values across the participants in the two treatment conditions. Such average effects are estimated at the second level of the multilevel model. The Level 2 model is:

Level 2

$$\begin{aligned}\alpha_i &= \zeta_{00} + \zeta_{01}\,\text{CONDITION}_i + r_{0i} \\ \beta_{1i} &= \zeta_{10} + \zeta_{11}\,\text{CONDITION}_i + r_{1i} \\ \beta_{2i} &= \zeta_{20} + \zeta_{21}\,\text{CONDITION}_i + r_{2i} \\ \beta_{3i} &= \zeta_{30} + \zeta_{31}\,\text{CONDITION}_i + r_{3i}\end{aligned} \tag{9.14}$$

where

CONDITION_i is an indicator variable coded 0 for the comparison group and 1 for the treatment group;
ζ_{00} is the mean of the α_i parameters for the comparison group;
ζ_{01} is the difference in the means of the α_i parameters between the comparison and treatment groups;
ζ_{10} is the mean of the β_{1i} parameters for the comparison group;
ζ_{11} is the difference in the means of the β_{1i} parameters between the comparison and treatment groups;
ζ_{20} is the mean of the β_{2i} parameters for the comparison group;
ζ_{21} is the difference in the means of the β_{2i} parameters between the comparison and treatment groups;
ζ_{30} is the mean of the β_{3i} parameters for the comparison group;
ζ_{31} is the difference in the means of the β_{3i} parameters between the comparison and treatment groups; and
The *r* terms are residuals (how much each parameter varies around its group mean).

(Note how Equation 9.14 is essentially the same as Equation 9.9 just with CONDITION_i substituted for SEX_i.) The estimates of ζ_{11} and ζ_{31} are estimates of the treatment effects (above and beyond any pseudo effects). The estimate of ζ_{11} reveals whether the comparison and treatment groups differ on average in the change in level that is due to the treatment. The estimate of ζ_{31} reveals whether the comparison and treatment groups differ on average in the change in slope that is due to the treatment. The estimates of ζ_{10} and ζ_{30} in Equation 9.14 are estimates of pseudo treatment effects in the comparison time series. The effects of the treatment in the experimental time series are estimated as differences above and beyond the pseudo effects in the comparison time series. Again, finding no pseudo effects would increase the credibility of the estimates of real treatment effects.

9.8.3 Caveats in Interpreting the Results of CITS Analyses

As noted at the start of Section 9.8, I would begin the analysis of data from the CITS design by plotting the data. As also noted, I would model the data from the treatment and comparisons time series separately using models such as those in Sections 9.4 and

9.5, allowing for treatment effects or pseudo treatment effects, respectively, in the treatment and comparison conditions. As I will explain shortly, I would be vigilant with subsequent analyses of the data if I found pseudo treatment effects in the comparison time series. Pseudo treatment effects might mean that effects (such as those due to history and instrumentation) are present in the comparison time series and hence could be a threat to internal validity when analyzing the data from the treatment time series. Pseudo treatment effects are evidenced by the β_1 and β_3 parameters in Equation 9.1 and the ζ_1 and ζ_3 parameters in Equation 9.8 when these models are fit to the data from the comparison time series. When performing the CITS analyses, pseudo treatment effects in the comparison time series are evidenced by the parameters β_2 and β_6 in Equation 9.12 and the parameters ζ_{10} and ζ_{30} in Equation 9.14. (Pseudo treatment effects could also arise because curvilinear trends in the data were not modeled correctly. This could raise doubts about whether curvilinear trends might also arise in the treatment time series that require proper modeling.)

If no pseudo effects are evident in the comparison time series and if, as a result, the researcher can be confident that effects due to shared threats to internal validity such as history and instrumentation are not present in the treatment time series, the researcher can be confident in interpreting the estimates of treatment effects in the CITS analyses in Section 9.8. But researchers need to be careful in interpreting treatment effects in CITS analyses in the presence of pseudo treatment effects, such as those due to shared history and instrumentation effects. The reason is that the effects of threats to internal validity might not be the same size in the treatment and comparison conditions. Consider an example. Suppose there is a shared history effect in both the treatment and comparison time series. Further, although there is a shared history effect in both the treatment and comparison time series, suppose the shared history effects are numerically different in the two time series. Then the simultaneous analyses using Equation 9.12 in Section 9.8.1.3 or Equations 9.13 and 9.14 in Section 9.8.2 would produce biased estimates of the treatment effect. This is because the difference in the sizes of the shared (but numerically different) history effects would be attributed to the effect of the treatment.

For example, consider comparing changes in traffic fatalities per capita due to an intervention instituted in state A while using state B as a comparison time series. Using per capita measures rather than total numbers of traffic fatalities might be considered a way to equate the two states with different population sizes and so ensure that shared history effects, for example, are of equal size. However, a per capita adjustment might not accomplish the task of equalization. Suppose the treatment is a change in speed limit, in the experimental state, from 65 to 55 on freeways, but there is no change in the speed limit in the comparison state. Further suppose there is a shared history effect in both states: airbags were introduced into new automobiles in both states at the same time the reduction in speed limit was introduced. If the residents in the experimental state were wealthier, on average, than the residents in the comparison state (so the residents in the experimental state bought more new vehicles with air bags), there could be

a larger reduction in traffic fatalities in the experimental state than in the comparison state from the shared history effects due to air bags, even if fatalities were measured per capita. As a result, the simultaneous analyses in Equations 9.12 in Section 9.8.1.3 and Equations 9.13 and 9.14 in Section 9.8.2 would incorrectly attribute the difference in history effects to an effect of the treatment.

Researchers cannot blindly assume that the estimates of treatment effects from the analyses of CITS data are free from bias due to shared threats to internal validity. Once a pseudo treatment effect is detected, the researcher must ask if the effects of shared threats to internal validity are the same size in the two time series and draw conclusions appropriately. If the effects of shared threats to internal validity are the same size in the two time series, the CITS analysis removes bias due to those threats. If the effects of shared threats to internal validity are different in size, however, the CITS analysis would not perfectly remove bias due to those threats. Instead, the researcher would have to argue that the treatment effect estimate was sufficiently large that it could not plausibly be explained in total as the result of shared threats to internal validity. For example, if a pseudo change in level due to a threat to internal validity in the comparison time series were estimated to be 4 and the change in level due to that shared threat to internal validity could be as large as 6 in the treatment time series, then the change in level in the treatment time series would have to be at least as large as 6 to be properly declared a treatment effect. In contrast, a blind analysis of the CITS data using Equation 9.12 or 9.13/9.14 would mistakenly conclude there was a treatment effect as long as the change in level in the treatment time series was simply greater than 4. The point here is that the CITS analyses are useful; it is just that the results must often be interpreted cautiously. To repeat, the estimates of treatment effects in the CITS analyses represent real treatment effects only if they are larger than can plausibly be explained by pseudo effects in the comparison time series. The blind application of the CITS analyses will not automatically provide treatment effect estimates that are free from the effects of pseudo effects.

In assessing whether the comparison time series exhibits a pseudo treatment effect, be sensitive to inappropriately accepting the null hypothesis. It is possible for pseudo treatment effects in the comparison time series to be large, even though they are not statistically significant. Judge the size of pseudo treatment effects based on the size of estimates of pseudo treatment effect parameters, in conjunction with accompanying confidence intervals, rather than solely on the results of statistical significance tests.

In general, the more similar the comparison time series is to the treatment time series, the greater confidence researchers can have in estimates of treatment effects in CITS designs. That is, the more similar the two time series are, the greater confidence the researcher can have that threats to internal validity have equal size effects in the two time series. In particular, it can increase confidence if the treatment and comparison conditions have equivalent pretreatment time series in both level and trend. A relatively recent innovation is the use of synthetic control time series (Abadie, Diamond,

& Hainmueller, 2015). When multiple comparison time series are available, synthetic control time series are created by weighting and aggregating the multiple comparison time series so that pretreatment observations are as similar as possible to the pretreatment observations in the treatment time series.

9.9 DESIGN SUPPLEMENTS III: COMPARATIVE ITS DESIGNS WITH MULTIPLE TREATMENTS

The basic CITS design can be supplemented by introducing multiple interventions rather than just one. One variant is what Shadish et al. (2002) call the switching replication design and is called a multiple-baseline (MBL) design in applied behavior analysis research (Kazdin, 2011). The design is diagrammed as:

NR: O_1 O_2 O_3 O_4 O_5 X O_6 O_7 O_8 O_9 O_{10} O_{11} O_{12} O_{13} O_{14} O_{15}

NR: O_1 O_2 O_3 O_4 O_5 O_6 O_7 O_8 O_9 O_{10} X O_{11} O_{12} O_{13} O_{14} O_{15}

There are now two experimental time series in which each receives the treatment but at different times. When the first treatment is introduced, an interruption should occur in the first time series but not in the second. While at the time of the second treatment implementation, an interruption should occur in the second time series but not in the first. Each time series serves as a comparison time series for the other, just at different times.

West et al. (1989) used a switching replication design (with an additional comparison time series) to assess the effects of a law strengthening penalties for driving under the influence of alcohol. The time series came from the cities of San Diego, Phoenix, and El Paso. Laws adding to the penalties for driving under the influence of alcohol were enacted in San Diego in January 1982 and in Phoenix in July 1982. No such laws were enacted in El Paso at any time. The results followed the expected pattern: large decreases in fatal traffic accidents in San Diego and Phoenix at the times the tougher laws were enacted in those cities, with no changes at those times in the comparison segments of the two time series (or in the El Paso time series).

The switching replication design can help rule out threats to internal validity, such as those due to history and instrumentation. For example, if history is to explain the predicted pattern of treatment results, history effects must occur in alternating fashion at specific points in time in each of the two time series. While such history effects may well be present, if the predicted outcome arises, it is relatively less plausible that a substantial history effect is present twice at just the right times and alternatively in the two time series than that two treatment effects are present. Hence, threats to internal validity can be less plausible as an explanation for the results in the switching replication CITS design than in the basic ITS design.

Alternatively, a crossover design involving two different treatments (X_A and X_B) could be implemented, which would be diagrammed as

NR: O_1 O_2 O_3 O_4 O_5 X_A O_6 O_7 O_8 O_9 O_{10} X_B O_{11} O_{12} O_{13} O_{14} O_{15}

NR: O_1 O_2 O_3 O_4 O_5 X_B O_6 O_7 O_8 O_9 O_{10} X_A O_{11} O_{12} O_{13} O_{14} O_{15}

In this design, the order of the two treatments is reversed in the two time series. Treatment X_A precedes X_B in one time series and follows X_B in the other time series. Again, the advantage of such designs compared to a basic ITS design is that the pattern of results due by the effects of the treatments in the complex CITS design is relatively difficult to explain plausibly by threats to internal validity. However, the researcher needs to be aware of order or carryover effects wherein a treatment can have different effects depending on whether it follows another treatment. The statistical analyses of complex CITS designs are generalizations of the analysis of basic CITS designs (Huitema, 2011).

9.10 SINGLE-CASE DESIGNS

With or without supplements, ITS designs are sometimes called **single-case designs** (SCD), especially when the time series of observations are short and the purpose is to assess behavioral interventions, such as in the field of applied behavioral analysis (Shadish, 2014b; Smith, 2012). In addition, a CITS design with switching replications (see Section 9.9) is sometimes called a **multiple-baseline (MBL) design,** especially with short time series and behavioral interventions. In SCD and MBL research, researchers often rely on analysis by visual inspection rather than statistical analysis as described so far (Kratochwill et al., 2010, 2013, 2014; Nugent, 2010). Indeed, Kratochwill et al. (2014, p. 93) note that "the vast majority of published single-case intervention research [such as in the field of applied behavior research] has incorporated visual analysis as the primary method of outcome evaluation."

Although visual analysis remains the primary means of data analysis for SCD and MBL designs, the use of appropriate statistical analyses is increasingly being encouraged, although their use is still contentious (Shadish, 2014a, 2014b; Smith, 2012). To a good extent, the proposed statistical methods are much the same as those used with ITS and CITS designs (especially hierarchical or multilevel models), as described above, with appropriate adaptations for limited amounts of data (Shadish et al., 2014). So as not to be redundant, I will focus here only on visual analysis.

When ITS designs are used in applied behavior analysis, a period of no treatment or removed treatment is denoted with an A and a period of treatment is denoted with a B, and both A and B periods are called **phases** (Kazdin, 2011). Phases labeled C, D, E, and so on designate additional, different treatments. The basic ITS design with a single introduction of the treatment is an AB design with two phases and a single

transition between baseline and treatment phases. A removed (also called a reversal or withdrawal) treatment design is an ABA design with three phases and two transitions: (1) from baseline to treatment and (2) from treatment to treatment removal. The repeated treatment design is an ABAB design, with four phases and three treatment transitions: (1) from baseline to the first introduction of the treatment, (2) from the first introduction of the treatment to treatment removal, and (3) from treatment removal to treatment reintroduction.

A multiple baseline design would have more than one treatment time series. Different participants could each contribute a time series of observations where each time series might be an ABA or an ABAB design. Or the same participant could contribute more than a single time series in different settings or with different outcome measures, where each time series might be an ABA or an ABAB design. The multiple time series in an MBL design are usually collected in parallel at the same time but with the treatment interventions being staggered to create a switching replication design.

Smith (2012) reviews six sets of standards for the visual analysis of data from SCDs. One of the most widely accepted sources of guidelines is the What Works Clearinghouse (WWC; Horner & Odom, 2014; Kratochwill et al., 2010, 2013; Smith, 2012; U.S. Department of Education, 2017). According to the most recent pilot standards established by the WWC, an SCD must have at least four phases (as in an ABAB design), with at least five data points of observation in each A and B phase to receive the rating of "Meets WWC Pilot SCD Standards Without Reservations." To receive the rating of "Meets WWC Pilot SCD Standards Without Reservations," MBL designs must contain at least six A and B phases (rather than just four as in an ABAB design), with at least five time points of observations per A and B phase. For example, an AB design repeated on at least three different participants with each A and B phase containing at least five time points could meet evidence standards without reservations. Note that with visual analysis, neither a basic ITS design (denoted AB) nor a basic CITS design (with an AB experimental time series and a simple A comparison time series) could meet WWC evidence standards without reservations no matter how many pretreatment or posttreatment observations were available.

According to the WWC, judging an effect to be present relies on visually assessing six patterns in the data: level, trend, variability, immediacy, overlap, and consistency. Level, trend, and variability are assessed within each A and B phase whereas immediacy, overlap, and consistency are assessed across phases. Ideally, the level and trend in the data within each phase would be easily discernible, and the data would exhibit little variability around the observed level and trend in each phase. In addition, differences across phases would appear immediately, with little overlap in levels between phases. Finally, the different phases would be consistent: each A phase would be similar to the other A phases, and each B phase would be similar to the other B phases. To conclude that a causal relation between treatment and outcome has been demonstrated in a single-case design, there would have to be at least three observed effects at three different transitions between A and B phases. In all cases, convincing visual analysis

requires treatment effects that are large enough to stand out dramatically against baseline variability.

9.11 STRENGTHS AND WEAKNESSES

The ITS design is fundamentally different from between-groups comparisons such as randomized experiments, nonequivalent group designs, and regression discontinuity designs. In those other three prototypical designs, the estimate of the treatment effect is derived primarily by comparing the performances of different groups of participants at the same point in time. In the ITS design, the estimate of the treatment effect is derived primarily by comparing performances of the same participant(s) at different points in time. Rather than using a group of participants to provide a counterfactual comparison, the ITS design uses a baseline of observations to provide its counterfactual comparison. This difference between the design types in their counterfactual comparisons is responsible for the relative advantages and disadvantages of the ITS design compared to the other designs.

The ITS design can be implemented with very few participants or units. Indeed, the design can be implemented with a single participant or unit. Therefore, the ITS design allows researchers to estimate treatment effects for a single participant or unit. This *N*-of-one feature has clear benefits in cost and ease of use compared to the other prototypical designs that need a much larger *N*. With either *N*-of-1 or small *N* and a short time lag between observations, the ITS design might allow researchers to quickly conduct a series of studies where each one builds on the other (Biglan, Ary, & Wagenaar, 2000). Such a series of studies can be especially important when a researcher wants to compare a large variety of interventions to see which ones work best or when a researcher needs to iteratively refine a treatment to be most effective.

Under the right conditions, the ITS design can also allow researchers to assess the temporal pattern of effects more easily than can often be done with the other designs. Another benefit is that the basic ITS design does not require the treatment to be withheld from any participants in the study. In between-groups comparisons (e.g., in randomized experiments, nonequivalent group designs, and regression discontinuity designs), a substantial proportion of participants do not receive the experimental treatment. Withholding the treatment from some participants can lead to resistance or resentment and make it more difficult to implement between-groups designs compared to ITS designs.

The relative advantages of the ITS design also come with potential disadvantages. To achieve the benefits of the ITS design, the researcher must obtain multiple observations on participants over time. In some research settings, collecting multiple observations over time may be relatively easy, especially with small-*N* designs, but it might not always be feasible. The ITS design is likely to produce less credible results than other designs either when the treatment has a gradual rather than an abrupt effect over time or when the time lag between treatment implementation and effect is delayed and

unknown. The problem is that gradual or delayed and unknown treatment effects allow for greater uncertainty about the effects of threats to internal validity and greater opportunity for threats to internal validity to arise.

Differences between design types produce differences in power and precision. Power and precision in between-groups comparisons depend on the number of participants and the heterogeneity among the outcomes of the participants. The more homogeneous the outcomes are within the treatment groups, the greater are power and precision. In contrast, power and precision in a basic ITS design depend mostly on the number of observations over time and the heterogeneity of those observations over time. The more homogeneous the observations over time, the greater are power and precision.

The ITS design is similar to the pretest–posttest design (see Chapter 6) in that they both estimate treatment effects by comparing participants before and after a treatment is introduced. The ITS design tends to produce more credible results, however, because it more readily rules out threats to internal validity, especially when the basic ITS design is supplemented with other design features. These other design features include withdrawing and/or repeating treatments and adding a comparison time series to the experimental time series. Each of these supplements can increase the number of comparisons that are being drawn and hence increase the number of opportunities for the treatment to evidence its effects in contrast to a comparison condition. These supplements increase the number of results that must be explained by threats to internal validity if the credibility of the estimates of the treatment effects is to be called into question. As the number of treatment comparisons increases, it becomes increasingly less plausible that they can be explained as being due to threats to internal validity rather than to the treatment (see also Chapters 11 and 12).

To assess the credibility of ITS designs, studies have drawn empirical comparisons between the results from ITS designs and the results from other high-quality designs, especially randomized experiments. For instance, Bloom, Michalopoulos, and Hill (2005) compared two classes of ITS designs, each of which included a comparison time series. In the benchmark class, the participants in the treatment and comparison conditions were assigned at random. In the other class of ITS design, the participants were not assigned at random to the treatment and comparison conditions. To be clear, the two classes of designs shared the same experimental time series but used different comparison time series. Cook and Wong (2008a, p. 148) drew the following conclusion about the Bloom et al. (2005) study: there was "little or no difference between the [randomized] control and [nonequivalent] comparison groups around the intervention and hence, there would be little causal bias" in the ITS design with the nonequivalent comparison group compared to the ITS design with the randomized comparison group. St. Clair, Cook, and Hallberg (2014) also compared a CITS design to a randomized experiment. The authors concluded (p. 324) that the ITS design with the nonequivalent comparison time series "provided valid causal estimates that hardly differed from the [randomized control trial] results." St. Clair et al. (2014) summarize the results from

four other studies. Somers et al. (2013) created a different comparison but with the same conclusion. They compared the results from an ITS design with a comparison group to the results from a well-implemented regression discontinuity design. They concluded that the ITS design produced valid estimates of treatment effects, as assessed by the regression discontinuity design.

Of course, we should not draw too strong a generalization based on the results of such a relatively small collection of studies because of their limited external validity. But these studies show that the ITS design is capable of producing results that are as credible as the results from randomized experiments and RD designs. Such a conclusion agrees with the general consensus about the ITS design: A well-implemented ITS design can produce highly credible estimates of treatment effects, especially when the ITS design is implemented with supplements such as reversed or repeated treatment interventions and/or comparison time series.

9.12 CONCLUSIONS

The ITS design is an enhancement of the pretest–posttest design and, as a result, produces more credible results than the pretest–posttest design. A well-implemented ITS design (especially with a comparison time series) can produce highly credible estimates of treatment effects. These estimates can be more credible than estimates from nonequivalent group designs and appear, at least under some conditions, to be similar in credibility to estimates from well-implemented randomized experiments and RD designs. Between-groups randomized experiments, nonequivalent group designs, and RD designs estimate treatment effects by drawing comparisons across different participants who receive different treatments. In contrast, the fundamental comparison in ITS designs is a comparison across times rather than across participants. Researchers should keep both types of comparisons in mind when designing studies to estimate treatment effects.

9.13 SUGGESTED READING

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science, 1*(1), 31–49.
—Provides examples and an accessible introduction to ITS designs in the context of community intervention research.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.
—The classic reference on ARMA models.

McCleary, R., McDowall, D., & Bartos, B. (2017). *Design and analysis of time series experiments*. New York: Oxford University Press.

—Also details ARMA analysis and is more readable than Box, Jenkins, and Reinsel (2008).

The following articles and books detail the analysis of panel time-series data, including data from CITS designs:

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering metrics: The path from cause to effect*. Princeton, NJ: Princeton University Press.

—Present difference-in-differences models for the analysis of CITS data.

Duncan, T. E., & Duncan, S. C. (2004a). A latent growth curve modeling approach to pooled interrupted time series analyses. *Journal of Psychopathology and Behavioral Assessment, 26*(4), 271–278.

Duncan, T. E., & Duncan, S. C. (2004b). An introduction to latent growth curve modeling. *Behavior Therapy, 35,* 333–363.

—Provide an introduction to the growth curve structural equation modeling approach to panel time-series data.

Kazdin A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford, UK: Oxford University Press.

Nugent, W. R. (2010). *Analyzing single system design data*. Oxford, UK: Oxford University Press.

—Detail the design and analysis of short time series in applied behavior analysis.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*(3), 385–405.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. Retrieved from *www.mdrc.org/publication/validity-and-precision-comparative-interrupted-time-series-design-and-difference*.

10

A Typology of Comparisons

No design is perfect. Each design has its own strengths and weaknesses. Each is susceptible to threats to validity, whether they are threats to internal, external, construct or statistical conclusion validity. In addition, each research circumstance imposes different constraints and demands. Designing a research study involves selecting from among the many possible design options so as to tailor the study to best fit the specific demands, including threats to validity, of the given research circumstances.

—REICHARDT (2009, p. 68)

It makes sense, when planning a study, to consider as complete a range of design options as possible.

—REICHARDT (2006, p. 7)

Overview

Preceding chapters have described research designs in which treatment effects are estimated by drawing comparisons across either participants or times. It is also possible to estimate treatment effects by drawing comparisons across settings and outcome measures. All four types of comparisons (across participants, times, settings, and outcome measures) can take the form of either randomized experiments or quasi-experiments.

10.1 INTRODUCTION

In the preceding chapters, I described a variety of designs from preexperimental designs to quasi-experimental designs to randomized experiments. I described the most common, prototypical forms of each of these designs—the designs most commonly used in practice. These are the designs most often described in methodological texts, but they do not cover the complete landscape of experimental and quasi-experimental designs. The time has come to explore the full terrain. First, however, I need to introduce the principle of parallelism and review the nature of the ideal comparison that defines a treatment effect.

10.2 THE PRINCIPLE OF PARALLELISM

Section 3.2 explained how an effect size is a function of the five size-of-effect factors of the cause, participant, time, setting, and outcome measure. The principle of parallelism states that if a design option is available for any one of the four size-of-effect factors of participant, time, setting, and outcome measure, the same design option is available as well for the other three size-of-effect factors (Reichardt, 2006). (The size-of-effect factor of the cause plays a unique role in designs and so is not part of the principle of parallelism, except in special cases such as labeling threats to validity; see Reichardt, 2006.) For example, if a treatment effect can be estimated by drawing a comparison across different participants, a treatment effect can also be estimated by drawing a comparison across different times, settings, or outcome measures. Similarly, if a comparison can be supplemented by adding a comparison involving different participants, a design can also be supplemented by adding a comparison involving different times, settings, or outcome measures.

Section 2.1 defined a treatment effect as the difference between what happened after a treatment was implemented and what would have happened if a comparison treatment had been implemented instead of the treatment, everything else having been the same. Section 2.2 labeled this counterfactual comparison the ideal comparison because it is impossible to obtain in practice. The ideal comparison is impossible to obtain in practice because everything else cannot be held the same in the comparison that is being drawn. Something else must vary along with the treatment conditions. This something else is either the participants, times, settings, or outcome measures—the four size-of-effect factors involved in the principle of parallelism. For example, it is possible to estimate a treatment effect by comparing what happens when the treatment and comparison conditions are given to two different groups of participants. But then the participants are not held constant; they differ across the two treatment conditions and so are confounded with the treatment. Similarly, it is possible to estimate a treatment effect by comparing what happens when the treatment and comparison treatment are given to the same participants at different times. But then the times are not held constant: they differ across the two treatment conditions and so are confounded with the treatment.

In many practical comparisons, more than one of the four size-of-effect factors of participant, time, setting, and outcome measure will vary, along with the treatment conditions. For example, consider a comparison where one group of participants is given the treatment and a different group of people receives the comparison condition. It is easy to see that the participants vary with the treatments. In this case, however, the settings will also differ with the treatments because the two groups of participants cannot share precisely the same environments. Or consider a pretest–posttest comparison or a basic interrupted time-series (ITS) design. In this case, time varies with the treatment conditions, but the participants necessarily vary as well because over time the participants get older and so are never identical at different times.

Even though more than one of the four size-of-effect factors of participant, time, setting, and outcome measure will always vary with the treatment conditions in any

comparison used to estimate an effect size, one of these factors usually varies most prominently with the treatment conditions. For example, time varies more prominently than any other size-of-effect factor in pretest–posttest comparisons and in basic ITS designs. Participants vary most prominently in nonequivalent group designs and in regression discontinuity designs. The size-of-effect factor that varies most prominently with the treatment conditions in a practical comparison is called the **prominent size-of-effect factor** in that comparison. Each of the four size-of-effect factors of participant, time, setting, and outcome measure can be the prominent size-of-effect factor. Moreover, each of these four prominent size-of-effect factors defines a type of comparison. In other words, there are four basic types of comparisons based on the four size-of-effect factors of participant, time, setting, and outcome measure: **comparisons across participants, comparisons across times, comparisons across settings,** and **comparisons across outcome measures.** That each of these factors defines a type of comparison is an example of the principle of parallelism.

In the next sections, I describe each of these four types of design in more detail. In each case, I give examples of comparisons that are randomized experiments: a randomized comparison across different participants, times, settings, and outcome measures. In subsequent sections, I show how each of these four types can also take the form of a quasi-experiment.

10.3 COMPARISONS ACROSS PARTICIPANTS

Comparisons across participants are arguably the most common type of comparison, especially in field settings. In these comparisons, participants are the size-of-effect factor that varies most prominently with the treatment conditions. A comparison across participants means that different participants are given different treatments and that the effect of the treatments is assessed by comparing the outcomes from the different participants. For example, educational programs are often assessed by delivering different treatments to different groups of children; manpower training programs are often assessed by delivering different treatments to different groups of adults who are out of work; and psychotherapy treatments are often assessed by delivering different therapeutic treatments to different clients. If the treatments were assigned to participants at random, the comparison across participants would be a randomized experiment. (Regression discontinuity designs and nonequivalent group designs are examples of comparisons across participants where the treatments are not assigned to participants at random.)

10.4 COMPARISONS ACROSS TIMES

Comparisons across times are comparisons in which times are the size-of-effect factor that varies most prominently with the treatment conditions. For example, consider assessing the effects of medication on a patient who suffers from migraines. The patient

might alternate, perhaps on a random schedule, taking the medication and an indistinguishable placebo each time he or she suffers a migraine. An hour after taking the medication or placebo, the patient would record the degree of headache pain suffered. The effect of the medication would then be assessed by comparing the degree of pain on days in which the medication was taken to the degree of pain on the days when the placebo was taken. Because different times receive different treatments in such a comparison, times vary along with the treatments, which means that times are confounded with the treatment conditions. Because the effect of the treatment is being assessed by comparing different times (rather than different participants), the comparison is a comparison across times.

If the medication and the placebo were indeed taken on a random schedule, the comparison across times would be a randomized experiment. Examples of random assignment of treatment conditions to times can be found in Cialdini, Reno, and Kallgren (1990); Edgington (1987); McLeod, Taylor, Cohen, and Cullen (1986); Shadish et al. (2002); and West and Graziano (2012). (Pretest–posttest designs and ITS designs are examples of comparisons across times where the treatments are not assigned to times at random.)

Note that a comparison across times could be implemented with a single participant as in the earlier example involving migraines. Or the comparison could be implemented with a group of participants. For example, rather than assessing the effects of the migraine medication on a single person, the effects of the medication could be assessed on a group of study participants, each using a comparison across times.

10.5 COMPARISONS ACROSS SETTINGS

Comparisons across settings are comparisons in which settings are the size-of-effect factor that varies most prominently with the treatment conditions. In a comparison across settings, treatments are assigned to different settings. For example, consider assessing the effects of adding traffic lights to street corners. In that case, traffic lights are added to some intersections, while other intersections are left without traffic lights. Traffic accidents are then recorded at each of the intersections. The effects of the traffic lights are assessed by comparing the accidents at the intersections with traffic lights to the accidents at the intersections without traffic lights. Because different settings receive different treatments in such a comparison, settings vary with the treatments, which means that settings are confounded with the treatment conditions. Because the effect of the treatment is being assessed by comparing different settings, the comparison is one across settings. If traffic lights were assigned to intersections at random, the comparison across settings would be a randomized experiment. (Because the units of analysis in such a study are traffic intersections, some readers have wondered if traffic intersections might be conceptualized as the participants in the study rather than as the settings. This perspective will not work, however. People, rather than traffic intersections, are the participants because the study is investigating the behavior of people at traffic

intersections, not the behavior of traffic intersections.) Examples of studies that assign treatment conditions to settings can be found in Goldstein, Cialdini, and Griskevicius (2008); Reding and Raphelson (1995); Reynolds and West (1987); Shadish et al. (2002); Sherman and Weisburd (1995); and West and Graziano (2012). For example, Reynolds and West (1987) compared purchases at different convenience stores, some of which implemented an advertising campaign and some of which did not.

10.6 COMPARISONS ACROSS OUTCOME MEASURES

In comparisons across outcome measures, outcome measures are the size-of-effect factor that varies most prominently with the treatment conditions. In this type of comparison, different treatments are assigned to different outcome measures. For example, consider a television show designed to teach the letters of the alphabet to preschoolers. Thirteen letters are chosen to be presented, one per television episode, for 13 weeks. A sample of children watches the 13 shows. At the end of the 13 shows, the children's knowledge of all 26 letters is measured. The effect of the television show is assessed by comparing the performance of the children on the 13 letters taught on the show to the performance of the children on the 13 letters not taught on the show. In this case, the performances on the 26 letters represent different outcome measures. Because different outcome measures receive different treatments in comparisons across outcome measures, outcome measures vary along with the treatments, which means outcome measures are confounded with the treatment conditions. If the 13 letters chosen to be taught on the television show were chosen from the 26 letters of the alphabet at random, the comparison across outcome measures would be a randomized experiment.

Note that within-subject studies conducted in psychological laboratories often involve randomized comparisons across outcome measures. For example, learning experiments are often conducted using different stimuli, which are presented in different treatment conditions. In such studies, the effect of the treatment is assessed by comparing how participants perform in response to the different stimuli. Thus, such studies involve comparisons across outcome measures.

A comparison across outcome measures can be implemented with a group of participants, as in the example given above. Or the comparison can be implemented with a single participant. For example, rather than assessing the effects of the television show on a group of children, the effect of the television show can be assessed on a single child using the same type of comparison across outcome measures.

10.7 WITHIN- AND BETWEEN-SUBJECT DESIGNS

The labels of within-subject and between-subject designs are more traditionally used in research methods and statistics texts than the labels I have given here. To some extent,

these traditional labels could be used instead of mine. For example, the label between-subject designs could be used instead of comparisons across participants. Also, the label within-subject designs (or repeated-measures designs) could apply to either comparisons across times or comparisons across outcome measures. The obvious drawback with the traditional labels is that they do not well distinguish among all four types of comparisons that I distinguish among here, and neither of the two traditional labels well applies to comparisons across settings.

10.8 A TYPOLOGY OF COMPARISONS

Table 10.1 presents a typology of comparisons. So far, I have presented four types of comparisons: comparisons across participants, comparisons across times, comparisons across settings, and comparisons across outcome measures. Each of the four rows in Table 10.1 corresponds to one of these four types of comparisons. That is, the first row of Table 10.1 is for comparisons across participants, the second row for comparisons across times, the third row for comparisons across settings, and the fourth row for comparisons across outcome measures. That is, the four rows of comparisons are distinguished by which of the four size-of-effect factors varies most prominently with the treatment.

The four rows in Table 10.1 are crossed with three ways in which the most prominent size-of-effect factor can vary with the treatment conditions. That is, three ways in which the most prominent size-of-effect factor can vary with the treatment are

TABLE 10.1. A Typology of Comparisons

Prominent size-of-effect factor	Assignment to treatment		
		Nonrandom (Quasi-Experiment)	
	Random	Explicit quantitative ordering	Nonequivalent assignment
Participants	Randomized comparison across participants	Regression discontinuity design	Nonequivalent group design
Times	Randomized comparison across times	Interrupted time-series design	Nonequivalent comparison across times
Settings	Randomized comparison across settings	Comparison based on a quantitative ordering of settings	Nonequivalent comparison across settings
Outcome measures	Randomized comparisons across outcome measures	Comparison based on a quantitative ordering of outcome measures	Nonequivalent comparison across outcome measures

represented by the three columns in the table. The fact that the four rows are crossed with the three columns embodies the principle of parallelism: whatever option is available for one size-of-effect factor is also available to the other size-of-effect factors (except the size-of-effect factor of the cause).

If the most prominent size-of-effect factor varies randomly with the treatment conditions, the comparison is a randomized experiment. The first column in Table 10.1 is for randomized experiments. As I have already shown in the preceding sections, each of the four types of comparisons defined by the most prominent size-of-effect factors (i.e., each of the rows in Table 10.1) can take the form of a randomized experiment.

If the most prominent size-of-effect factor varies nonrandomly with the treatment conditions, the comparison is a quasi-experiment. Two types of quasi-experiments are distinguished, as shown in the second and third columns of Table 10.1. The second column of the table contains comparisons where the most prominent size-of-effect factor is assigned to treatment conditions based on an explicit quantitative ordering. In the most common form of that type of assignment, the units of the most prominent size-of-effect factor are ordered quantitatively and assigned to treatment conditions based on a single cutoff score on that quantitative dimension. Each of the four types of comparisons defined by the most prominent size-of-effect factor (i.e., each of the rows in Table 10.1) can take the form of a "cutoff score" design.

The third column in Table 10.1 is for quasi-experimental comparisons where the units of the most prominent size-of-effect factor are assigned to treatments neither at random nor according to a cutoff score on an explicit quantitative ordering. Each of the four types of comparisons defined by the most prominent size-of-effect factor (i.e., each of the rows in Table 10.1) can have this type of assignment to treatment conditions.

10.9 RANDOM ASSIGNMENT TO TREATMENT CONDITIONS

As noted, the first column of Table 10.1 is for randomized experiments. Each of the four types of comparisons defined by the most prominent size-of-effect factor (i.e., the comparisons in each of the four rows in the table) can take the form of a randomized experiment.

I have already given examples of each of the four types of randomized experiments, but let me repeat to emphasize the point. (1) A comparison across participants can take the form of a randomized experiment; that is, different participants can be assigned at random to different treatments. (2) A comparison across times can take the form of a randomized experiment; for example, a patient can take migraine medications or a placebo on randomly assigned days. (3) A comparison across settings can take the form of a randomized experiment; for example, traffic intersections can be assigned randomly to have a traffic light. (4) A comparison across outcome measures can take the form of a randomized experiment; for example, letters of the alphabet can be assigned randomly to be taught on a television program.

10.10 ASSIGNMENT TO TREATMENT CONDITIONS BASED ON AN EXPLICIT QUANTITATIVE ORDERING

The second column in Table 10.1 contains quasi-experimental comparisons where assignment to treatment conditions is based on an explicit quantitative ordering of the items of the most prominent size-of-effect factor. The most common form of such **assignment based on an explicit quantitative ordering** is when items of the most prominent size-of-effect factor are ordered along a quantitative dimension and assigned to treatment conditions based on a cutoff score on that dimension. That is, all items that have scores above the cutoff score would be assigned to one treatment condition, while all items that have scores below the cutoff score would be assigned to the other treatment condition.

Comparisons based on a quantitative ordering of participants appear in the first row and second column of Table 10.1. The regression discontinuity (RD) design is a comparison based on the quantitative ordering of participants. As described in Chapter 8, the RD design orders participants on a quantitative dimension, such as a measure of need or merit, and assigns them to treatments based on a cutoff score. For example, educational programs could be assigned to children based on their scores on a measure of academic ability; or manpower training programs could be assigned to adults based on the length of time they have been out of work; or psychotherapy treatments could be assigned to patients based on a measure of severity of depression. As described in Chapter 8, the effect of the treatment is estimated by regressing an outcome measure onto the quantitative assignment variable in each group of participants. A discontinuity in these regressions at the cutoff point is taken as evidence of a treatment effect.

Comparisons based on a quantitative ordering of times appear in the second row and second column of Table 10.1. The interrupted time series design is a comparison based on the quantitative ordering of times. As described in Chapter 9, the ITS design orders times chronologically (which is an order on a quantitative dimension) and assigns times to treatment conditions based on a cutoff score. For example, days could be ordered chronologically and migraine medications or placebos could be assigned before and after a cutoff time. As described in the chapter on the interrupted time series design, the effect of the treatment is estimated by plotting the trend in the outcome measure over time both before and after the cutoff time. A discontinuity in these trends at the cutoff point is taken as evidence of a treatment effect. Hopefully, the similarity to the regression discontinuity design is clear (Marcantonio & Cook, 1994).

Comparisons based on a quantitative ordering of settings appear in the third row and second column of Table 10.1. In such designs, settings are ordered along a quantitative dimension and assigned to treatments based on a cutoff score on that dimension. For example, in assessing the effect of adding traffic lights to intersections, intersections could be ordered based on their traffic volume or frequency of accidents, and traffic lights could be added to those intersections that were busiest or had the most accidents. An estimate of a treatment effect would be derived in a comparison based on

a quantitative ordering of settings in a parallel fashion to the way a treatment effect is estimated in an RD or ITS design. That is, an outcome measure (e.g., number of traffic accidents) would be regressed onto the quantitative assignment variable (e.g., traffic volume) in each treatment condition, and a discontinuity in the regression lines at the cutoff score would be taken as evidence of a treatment effect.

Comparisons based on a quantitative ordering of outcome measures appear in the fourth row and second column of Table 10.1. In such designs, outcome measures are ordered along a quantitative dimension and assigned to treatments based on a cutoff score on that dimension. For example, in assessing the effects of a television show on teaching letters of the alphabet, the letters of the alphabet could be ordered, for example, on the frequency with which they are used in the English language. Then the television show could teach the letters that appear most frequently. An estimate of a treatment effect would be derived in a comparison based on a quantitative ordering of outcome measures in parallel fashion to the way a treatment effect is estimated in an RD or ITS design. That is, an outcome measure (e.g., knowledge of the letters of the alphabet) would be regressed onto the quantitative assignment variable (e.g., letters of the alphabet ordered by frequency of appearance in the English language) in each treatment condition, and a discontinuity in the regression lines at the cutoff score would be taken as evidence of a treatment effect.

10.11 NONEQUIVALENT ASSIGNMENT TO TREATMENT CONDITIONS

The third column in Table 10.1 contains quasi-experimental comparisons where assignment to treatment conditions is neither random nor based on a quantitative ordering of the items of the most prominent size-of-effect factor. Such assignment to treatment conditions is called **nonequivalent assignment to treatment conditions.** Comparisons across participants can be based on nonequivalent assignment to treatment conditions. Such comparisons are represented in the first row and third column of Table 10.1 and are called nonequivalent comparisons across participants. The nonequivalent group design is a nonequivalent comparison across participants. Nonequivalent assignment arises, for example, when people self-select the treatment they receive. For example, adults out of work could choose to enroll in manpower training programs based on their nonquantitatively assessed motivation to work. Or patients could seek psychotherapy treatment based on the nonquantitatively assessed degree of their symptoms. Nonequivalent assignment also arises when administrators or other third parties assign treatments based on nonquantitatively assessed characteristics. For example, educational programs could be assigned nonequivalently to children if parents enroll their children in the program based on their nonquantitatively assessed desires and inclinations. To estimate the effect of a treatment in a nonequivalent comparison across participants, the researcher must take account of the effects of the nonrandom selection

differences between participants in the different treatment conditions. Methods for such a task were described in Chapter 7.

Comparisons across times can be based on nonequivalent assignment to treatment conditions and are represented in the second row and third column of Table 10.1. These comparisons are called nonequivalent comparisons across times. For example, a patient suffering from migraines might take medication based on the nonquantitatively assessed severity of the person's headache on a given day. To estimate the effect of a treatment in a nonequivalent comparison across times, the researcher must take account of the effects of the nonrandom selection differences between times in the different treatment conditions. This task could be accomplished using methods like those described in Chapter 7 on the nonequivalent group design.

Comparisons across settings can be based on the nonequivalent assignment to treatments. Such comparisons are represented in the third row and third column of Table 10.1. These comparisons are called nonequivalent comparisons across settings. For example, traffic lights could be assigned to street corners based on complaints of local residents or the influence of local politicians. (If the frequency of complaints or the degree of influence could be quantified, the comparison could become a quantitative assignment comparison. If the frequency of complaints or the degree of influence is neither random nor quantified, however, the comparison is a nonequivalent setting comparison.) To estimate the effect of a treatment in a nonequivalent comparison across settings, the researcher must take account of the effects of the nonrandom selection differences between settings in the different treatment conditions. This task could be accomplished using methods like those described in Chapter 7 on the nonequivalent group design.

Comparisons across outcome measures can be based on nonequivalent assignment to treatment conditions. Such comparisons are represented in the fourth row and third column of Table 10.1. These comparisons are called nonequivalent comparisons across outcome measures. For example, letters of the alphabet could be chosen to be taught on a television show based on the whim of the show's producer. To estimate the effect of a treatment in a nonequivalent comparison across outcome measures, the researcher must take account of the effects of nonrandom selection differences between outcome measures in the different treatment conditions. This task could be accomplished using methods like those described in Chapter 7 on the nonequivalent group design.

10.12 CREDIBILITY AND EASE OF IMPLEMENTATION

The 12 types of comparisons displayed in Table 10.1 do not generally produce results that are equally credible. Credibility tends to decrease as you move across the comparisons in the table from left to right. That is, randomized experiments tend to produce more credible results than quantitative assignment comparisons, which tend to produce more credible results than comparisons based on nonequivalent assignment.

Differences in credibility depend largely on the ability to take account of the threat to internal validity due to selection differences. Selection differences are present in all the comparisons in Table 10.1. In comparisons across participants, selection differences are differences between the participants in the different treatment conditions. In comparisons across times, selection differences are differences between the times in the different treatment conditions; in comparisons across settings, selection differences are differences between the settings in the different treatment conditions; and in comparisons across outcome measures, selection differences are differences between the outcome measures in the different treatment conditions.

To continue the discussion of differences in credibility, consider how different participants are assigned to the different treatment conditions in all comparisons across participants (i.e., in the first row in Table 10.1). Selection differences between participants are present in randomized comparisons across participants, just as they are present in comparisons with the quantitative assignment of participants and in nonequivalent comparisons across participants. Differences in credibility arise not from the presence or absence of selection differences, but rather because of the nature of the selection differences that are present in the different types of comparisons and the ease with which selection differences can be taken into account. In randomized comparisons, selection differences are random (see Chapter 4 on randomized experiments). The effects of random selection differences can be easily taken into account using simple statistical methods such as an ANOVA. In comparisons with quantitative assignment to treatment conditions, the effects of selection differences are taken into account by fitting regression surfaces to data from each treatment condition and drawing comparisons across these regression surfaces (see Chapter 8 on the RD design). Such fitting and comparing of regression surfaces involve assumptions that are not required in randomized experiments. Being required to make more assumptions, makes the estimates of treatment effects more tenuous in comparisons with the quantitative assignment to treatment conditions than in randomized comparisons. Taking account of selection differences in nonequivalent comparisons across treatment conditions requires even more assumptions which lead to even more tenuous treatment effect estimates (see Chapter 7 on the nonequivalent group design). The same differences in assumptions and credibility hold for comparisons across times, settings, and outcome measures and not just for comparisons across participants.

However, which comparisons produce the most credible results depends not just on assumptions but also on the research circumstances. For example, randomized experiments are most credible in theory but, as noted in Chapter 4 on randomized experiments, they can degenerate into quasi-experiments. A degenerated (i.e., broken) randomized experiment can still produce more credible results than a quasi-experiment, but that will not always be the case. Circumstances can arise where a degenerated randomized experiment produces less credible results than quasi-experiments, especially well-implemented quasi-experiments involving quantitative assignment to treatment conditions. In theory, however, if everything else is the same

(which, of course, it never is in practice), credibility tends to decrease as you move from left to right in Table 10.1.

In contrast, the ease of implementation is just the opposite: it tends to increase in moving from left to right in Table 10.1. Nonequivalent comparisons tend to be easier to implement than comparisons with quantitative assignment to treatment conditions, which tend to be easier to implement than randomized experiments.

10.13 THE MOST COMMONLY USED COMPARISONS

Some comparisons in Table 10.1 are used in practice far more often than are others. Social psychologists, educational researchers, economists, and sociologists tend to use comparisons across participants more than other types of comparisons. In contrast, laboratory studies conducted by cognitive psychologists might well employ randomized comparisons across outcome measures (i.e., within-subject designs) as much as comparisons across participants (i.e., between-subject designs). Also, research in behavior modification uses comparisons over time (in the form of ITS designs) on single subjects more than any other comparisons. Agricultural research often uses randomized comparisons across settings. Program evaluation uses both randomized experiments (especially randomized experiments across participants) as well as quasi-experiments. Accordingly, textbooks on research methods focus differentially on research designs, depending on the substantive area of concern in the textbook. In many substantive areas, texts on statistical analysis emphasize randomized comparisons far more than quasi-experimental comparisons, and they especially tend to emphasize randomized comparisons across participants rather than randomized comparisons across times, settings, or outcome measures.

Among quasi-experiments, the nonequivalent group design, the RD design, and the ITS design are used most often, which is why I have devoted separate chapters to these designs. Popular texts on quasi-experimentation (Judd & Kenny, 1981; Mohr, 1995; Shadish et al., 2002) tend to ignore other types of comparisons, such as comparisons across settings or outcome measures. Readers of these texts likely come away without awareness of a good number of possible comparisons. For example, when texts on quasi-experimentation discuss randomized experiments, they tend to ignore randomized designs except for randomized comparisons across participants. Interestingly, the text that made quasi-experiments famous, Campbell and Stanley (1966), discusses randomized comparisons across times and outcome measures but labels them quasi-experiments rather than randomized experiments. For example, in Campbell and Stanley (1966) randomized comparisons across times and across outcome measures are called quasi-experiments under the labels of "equivalent time samples design" and "equivalent materials design," respectively.

Although only some types of comparisons tend not to be addressed in texts, researchers are well advised to consider the entire typology of designs in Table 10.1

in order to choose among the broadest range of comparisons possible. Until we have better theories of how to fit research designs to research circumstances, researchers should consider how each of the designs in Table 10.1 might well fulfill their research needs. Considering only some designs in the table means researchers may be overlooking designs that can be well tailored to the specific research setting. Here is an example. The original external evaluations of *Sesame Street* in the early 1970s used numerous designs, including randomized comparisons across participants and nonequivalent group designs (Ball & Bogatz, 1970: Cook et al., 1975). But the producers of the show could have easily supplemented their research with randomized comparisons across outcome measures. All that would have been needed was for the television show to select a random sample of letters to teach during the first year of production and compare viewers' knowledge of those letters to the letters not taught during the first year. The point is that some of the best designs might be overlooked if researchers do not consider all possible designs in Table 10.1. A complete theory of experimentation should explicate the full panoply of designs in Table 10.1.

10.14 CONCLUSIONS

Comparisons can be drawn across all four size-of-effect factors of participants, times, settings, and outcome measures. That is, a practical comparison, as opposed to the ideal comparison that defines a treatment effect, can be used to estimate the effect of a treatment by varying treatment conditions across participants, times, settings or outcome measures. Treatments can be assigned to either participants, times, settings, or outcome measures in three ways: at random, based on a quantitative ordering, or neither at random nor based on a quantitative ordering (i.e., nonequivalent assignment). Crossing the four size-of-effect factors with the three methods of assignment creates the four-by-three layout in Table 10.1. When designing a research study, choosing from among all the options in Table 10.1 can produce more credible results than choosing among only a limited number of prototypical designs.

10.15 SUGGESTED READING

Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods, 11,* 1–18.
—Goes into further detail about the principle of parallelism and how it applies to research designs.

11

Methods of Design Elaboration

Quasi-experimental designs require particular attention to the achievement of appropriate control groups, since the absence of random assignment raises serious dangers of noncomparability of groups and consequent uninterpretability of results. Many studies could be improved by the use of multiple control groups; the more different kinds of control groups (or control observations), the greater the number of rival hypotheses that can be rendered implausible, and the stronger the case for the causal relationship the experimenter has in mind.

—Webb and Ellsworth (1975; quoted in Roos, Roos, & Henteleff, 1978, p. 504)

It might be possible to include in the study response measurements or supplementary observations for which alternative hypotheses give different predictions. In this way, ingenuity and hard work can produce further relevant data to assist the final judgment.

—Cochran (1972/2015, p. 136)

The best quasi-experiments are not designed by simply picking a design from a list. Rather, they are created thoughtfully by combining many design features together to create designs that may often be more complex to execute both logically and logistically but that return that investment in better causal inference.

—Shadish and Luellen (2006, pp. 548–549)

Overview

Researchers can use one of three methods of design elaboration (which are special types of design supplements) to take account of threats to internal validity. In design elaboration, an estimate of a treatment effect is combined with a second estimate to disentangle the effects of a threat to internal validity from the effect of the treatment. In accord with the principle of parallelism (see Section 10.2), the two estimates in design elaboration can be derived from different participants, times, settings, or outcome measures.

11.1 INTRODUCTION

In other chapters, I have shown how including an additional comparison in a design can help rule out threats to internal validity. For example, adding a comparison time

series in an interrupted time series design can help rule out threats to internal validity, such as history effects and instrumentation (Section 9.8). In the present chapter, I describe three ways to add a comparison to a design to rule out threats to internal validity. I call these three ways of adding a comparison the three **methods of design elaboration** (Reichardt, 2000, 2006; Reichardt & Gollob, 1989) which are special forms of design supplements. When a comparison is added to rule out a threat to internal validity, it operates by way of one of these three methods of design elaboration. It behooves researchers to understand all three methods so that they can choose the method or methods that best fit their research circumstances.

11.2 THREE METHODS OF DESIGN ELABORATION

In the three methods of design elaboration, threats to internal validity are addressed by combining two estimates. When a threat to internal validity biases an estimate of a treatment effect, it means the estimate (which I will call the original or first estimate) equals the treatment effect, plus a bias due to the threat to internal validity. This can be represented schematically as

$$\text{Estimate}_1 = (\text{Treatment Effect}) + \text{Bias}$$

For example, when history is a threat to internal validity, a history effect can introduce bias in the estimate of the treatment effect that will either inflate or deflate the estimate.

In design elaboration, a second (or additional) estimate is obtained, which allows the treatment effect to be disentangled from the bias. There are three methods of design elaboration; these three methods combine two estimates to disentangle the treatment effect and the bias in the first estimate, and are distinguished by the relationship between the first and second estimates. These three methods are explicated in the following sections.

11.2.1 The Estimate-and-Subtract Method of Design Elaboration

In the **estimate-and-subtract method of design elaboration** (Reichardt & Gollob, 1989), the two estimates are

$$\begin{aligned} \text{Estimate}_1 &= (\text{Treatment Effect}) + \text{Bias} \\ \text{Estimate}_2 &= \qquad\qquad\qquad\quad \text{Bias} \end{aligned}$$

where the second estimate is simply an estimate of the bias in the first estimate. By subtracting the second estimate from the first, the researcher obtains an estimate of the treatment effect free from the bias:

$$\text{Estimate}_1 - \text{Estimate}_2 = \text{Treatment Effect}$$

Or even more simply, if the second estimate is equal to zero, it shows the bias is zero in both estimates, and so the first estimate equals the treatment effect without any bias. For example, consider a basic ITS design:

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad O_5 \quad O_6 \quad X \quad O_7 \quad O_8 \quad O_9 \quad O_{10} \quad O_{11} \quad O_{12}$$

This design produces an estimate of a treatment effect, but the estimate is susceptible to bias due to a history effect:

$$\text{Estimate}_1 = (\text{Treatment Effect}) + (\text{Effect of History})$$

Adding a comparison time series can remove the bias. With the comparison time series based on a nonequivalent group of participants, the design becomes

$$\begin{array}{lcccccccccccc} \text{NR:} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & X & O_7 & O_8 & O_9 & O_{10} & O_{11} & O_{12} \\ \hline \text{NR:} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & & O_7 & O_8 & O_9 & O_{10} & O_{11} & O_{12} \end{array}$$

The comparison time series that is obtained is free from the effects of the treatment but shares the same effects of history. That is, the second (comparison) time series would produce an estimate:

$$\text{Estimate}_2 = \text{Effect of History}$$

Subtracting the second estimate from the first produces an estimate of the treatment effect free from the bias:

$$\text{Estimate}_1 - \text{Estimate}_2 = \text{Treatment Effect}$$

Alternatively, by finding no pseudo interruption in the comparison time series, the researcher would conclude that there were no effects of history.

Rosenbaum (2017) provides another example of the estimate-and-subtract method of design elaboration. His example involves the effects of changes in economic incentives on weeks worked following an injury. The difference before and after the economic incentive that was implemented provided the first estimate. This estimate could have been biased by changes from before to after in background economic conditions. The second estimate was derived from a corresponding before–after comparison in a nonequivalent group of participants where no treatment (i.e., no change in economic incentives) was implemented but the same background economic forces were operating. That the second estimate revealed no differences due to changes in background economic

conditions supported the conclusion that the first estimate was free from the effects of bias due to changes in background economic conditions.

Rather than deriving the second estimate from a different time series of observations or from a different before–after comparison of a nonequivalent group of participants, as in the preceding two examples, the second estimate could be derived from "auxiliary" information involving direct measurements of potential biases. For example, in Connecticut's crackdown on speeding, history could have been a threat to internal validity (Campbell & Ross, 1968). Perhaps a gruesome accident was widely reported in the news and made people drive more carefully at the same time that the crackdown was implemented. If so, the gruesome accident rather than the crackdown on speeding might have been responsible for any reduction in traffic fatalities. To assess the plausibility of such an alternative hypothesis, a researcher could check newspapers for reports of accidents or ask drivers about their awareness of such events. The results of such an investigation could serve as a second estimate showing that no biases due to such effects were present. Or perhaps a reduction in traffic fatalities resulted because safety features (such as seat belts) were introduced at the same time as the crackdown on speeding. Again, a researcher could check the plausibility of that alternative hypothesis through focused auxiliary measurements that would again serve the function of a second estimate.

Rosenbaum (2017) provides examples of the estimate-and-subtract method of design elaboration under the label of "counterparts." He also notes that the method is sometimes called a method of difference in differences. That is, the method assesses the difference between the first and second estimates where each of these estimates is itself a difference.

11.2.2 The Vary-the-Size-of-the-Treatment-Effect Method of Design Elaboration

The **vary-the-size-of-the-treatment-effect method of design elaboration** is a generalization of the estimate-and-subtract method of design elaboration. In the vary-the-size-of-the-treatment-effect method, the second estimate equals a different amount of the treatment effect but the same amount of bias due to a threat to validity. Taken together, the two estimates are

$$\text{Estimate}_1 = \quad (\text{Treatment Effect}) + \text{Bias}$$
$$\text{Estimate}_2 = (\Delta \times \text{Treatment Effect}) + \text{Bias}$$

where Δ is any number not equal to 1 and "$\Delta \times$ Treatment Effect" means Δ times the treatment effect. A difference between the two estimates in the vary-the-size-of-the-treatment-effect method suggests that the treatment has an effect. Note that if Δ equals 0, the vary-the-size-of-the-treatment-effect method of design elaboration reduces to the

estimate-and-subtract method. In other words, the estimate-and-subtract method is a special case of the vary-the-size-of-the-treatment-effect method.

An example of the vary-the-size-of-the-treatment-effect method comes from Leibowitz and Kim (1992; Azar, 1994). Leibowitz and Kim estimated the effect of galanin, a substance that occurs naturally in rats, on weight gain in rats. In this study, some rats were injected with an additional amount of galanin and compared to a noninjected group of rats. A comparison of the two groups revealed that the additional galanin increased weight gain as expected. But because galanin was administered by injection while the comparison group received no injection, the effect of the injection was confounded with the treatment. That is, the first estimate was equal to

$$\text{Estimate}_1 = (\text{Treatment Effect of Galanin}) + (\text{Effect of Injection})$$

The effect of the injection was then disentangled from that of the treatment by adding a second comparison. In the second comparison, a substance that blocks the uptake of naturally occurring galanin was injected into a group of rats and was shown to decrease weight compared to a noninjected comparison group. In schematic form, the second estimate was:

$$\text{Estimate}_2 = (-1 \times \text{Treatment Effect of Galanin}) + (\text{Effect of Injection})$$

That the two estimates differed in direction reveals that the treatment effect is nonzero, despite the shared bias due to the injection.

For another example, consider the standard method of controlling for the biasing effects of yea saying (i.e., acquiescence bias). Yea saying is the tendency for participants in a study to respond positively to any question regardless of content. A traditional method to control for yea saying is to word half the items in a questionnaire, so that agreeing corresponds to a high level of the trait while the other half of the items are worded so that agreeing corresponds to a low level of the trait. The expected treatment effect is in opposite directions in the comparison of the two halves of the questionnaire, while the effect of yea saying is in the same direction. Subtracting one estimate from the other provides an estimate of the treatment effect unencumbered by the effects of yea saying.

Sections 7.10.5 and 9.7.1 explain how designs can be made more complex by adding comparisons involving reversed or removed treatments. Adding such comparisons can rule out shared threats to validity using the vary-the-size-of-the-treatment-effect method.

Note that the vary-the-size-of-the-treatment-effect method can be implemented in a variety of ways. The second estimate can be obtained by giving different amounts of the treatment to different groups of participants. The example involving the effects of galanin was such a case. The first estimate was obtained with injections of galanin in one group of rats. The second estimate was obtained with injections of a drug that

blocked the uptake of galanin in a second group of rats. Alternatively, a second estimate could be obtained by giving the same amount of treatment to different participants. An example comes from Glass (1988) in an assessment of the effects of desegregation (via mandated busing) on white flight from inner cities to suburbs. Busing was equally mandated for the wealthy and the poor but its effects were presumed to be larger for the wealthy because they could more easily afford to move to the suburbs to avoid busing. In this way, comparing the effects of busing among the wealthy and poor compared two different effects of the same treatment, which are presumed to share the same biasing effects of history. That different effects were evident for the wealthy and poor revealed that the treatment had an effect above and beyond any shared effects of history.

Here is another distinction between types of second estimates. The second estimate could be created by either adding data or disaggregating data. In the study of the effects of galanin, the second estimate was created by collecting additional data. In the study of busing and white flight, the second estimate was created by disaggregating data already at hand—that is, by dividing the available sample into wealthy and poor subsamples.

11.2.3 The Vary-the-Size-of-the-Bias Method of Design Elaboration

In the **vary-the-size-of-the-bias method of design elaboration,** the second estimate has the same amount of treatment effect but a different amount of bias. The two estimates are:

$$\text{Estimate}_1 = (\text{Treatment Effect}) + \quad (\text{Bias})$$
$$\text{Estimate}_2 = (\text{Treatment Effect}) + (\Delta \times \text{Bias})$$

where Δ differs from the value of 1 and "$\Delta \times$ Bias" means Δ times Bias. If the two estimates turn out to be the same size, the bias is shown to be zero.

A study by Minton (1975) cited in both Cook and Campbell (1979) and Shadish et al. (2002) provides an example. Minton was interested in the effects of the television show *Sesame Street* on intellectual development. To estimate the effect, she used a cohort design in which she compared younger siblings (who had watched the show) to older siblings who had not watched the show when they were the same age because *Sesame Street* had not yet been on the air. In this way, the two siblings were matched for age. This is a little complex, so let me explain in more detail. *Sesame Street* first aired in 1969. An older sibling who was 4 years old in 1968 was matched with a younger sibling who was 4 years old in 1970. The older sibling's performance in 1968 could not have been affected by the show because it had not aired yet. In contrast, the younger sibling's performance in 1970 could have been influenced by the show because it had aired already. Because the siblings were matched on age, age could not bias the comparison. But the siblings were necessarily not matched on birth order. It is well known (Zajonc & Markus, 1975) that older-born siblings tend to perform better than later-born siblings

on cognitive tasks. Cook and Campbell (1979) proposed using the vary-the-size-of-the-bias method to control for this bias. The first of the two estimates was based on a comparison of older siblings who were first born with younger siblings who were second born. The second estimate was based on a comparison of older siblings who were second born with younger siblings who were third born. While the size of the treatment effect should be the same in these two estimates, the size of the bias due to birth order should differ substantially because the effect of birth order is substantially larger for first-to-second born siblings than from second-to-third born siblings. Finding the two estimates to be the same size ruled out birth order as a plausible bias.

As another example of the vary-the-size-of-the-bias method of design elaboration consider controlling for the effects of social desirability, where participants respond to items on a questionnaire in ways they believe reflect the most socially desirable response. Divide the items on the questionnaire into two parts where the effect of social desirability is large in one part and small in the other. Then to the extent that the two parts produce estimates of effects that are the same size, bias due to social desirability is rendered implausible. Similar methods could be used to address potential biases due to experimenter and participant expectancies, volunteer effects, evaluation apprehension, and hypothesis guessing among other threats to validity. That is, a researcher could compare an estimate where the biases were presumed to be large to an estimate where the biases where presumed to be relatively small, but where the effect of the treatment was likely to be the same in both estimates. In using the vary-the-size-of-the-bias method in such ways, the researcher can rely on variation in the size of the bias that occurs naturally, or the researcher can induce variations purposefully.

11.3 THE FOUR SIZE-OF-EFFECT FACTORS AS SOURCES FOR THE TWO ESTIMATES IN DESIGN ELABORATION

As explained in the preceding sections, each of the three methods of design elaboration involves two estimates: a first and a second estimate. The two estimates come from different sources where the sources are defined by the four size-of-effect factors. That is, in keeping with the principle of parallelism, each of the four size-of-effect factors of participants, settings, times, and outcome measures can be a source for the two estimates in design elaboration. For example, in the methods of design elaboration, the first estimate could come from one set of participants, and the second estimate could come from a second set of participants; the first estimate could be collected at a given time, and the second estimate could be collected at a different time; the first estimate could come from one setting, and the second estimate could come from a different setting; or the first estimate could come from one outcome measure and the second from a different outcome measure. The present section explains how each of the four size-of-effect factors (of participants, times, settings, and outcome measures) can provide

the two estimates in design elaboration. By considering each of the four possibilities, a researcher can choose the option best suited to the research circumstances.

The four sources of the two estimates (participants, times, settings, and outcome measures) are crossed with the three methods of design elaboration. For simplicity, all the examples that follow are based on the estimate-and-subtract method of design elaboration. The same logic applies to the other two methods of design elaboration.

11.3.1 Different Participants

In the methods of design elaboration, the first estimate could come from one set of participants, and the second could come from a different set of participants. Many ITS designs that include a comparison time series are of this nature—that is, where the experimental and comparison time series are derived from different sets of participants. As an example, Wagenaar (1981, 1986) estimated the effects of a change in the legal drinking age in Michigan that increased the legal drinking age from 18 to 21 years of age. Wagenaar assessed the effects of this change on traffic accidents using a comparative ITS design with an experimental time series and two comparison time series of data from different participants. The experimental time series consisted of data from 18- to 20-year-old drivers. There was a decrease in traffic accidents in this time series corresponding to the change in the law. Such a drop would be expected if the law were effective in reducing drinking among those ages 18–20 years as the law intended. However, it is possible that the decrease in accidents was due to a history effect that occurred at the same time as the drinking age law was passed. Perhaps a different law was passed at the same time requiring the wearing of seat belts, or perhaps there was a police crackdown on speeding, or perhaps there was a widely published gruesome accident that caused people to drive more carefully. Wagenaar used the two comparison time series to rule out such history effects. One comparison series was traffic accidents involving drivers younger than 18, and another was drivers who were 21 and older. Neither of these time series should have shown an effect of the change in drinking age (because the law should only have influenced those ages 18–20), but both should have been susceptible to the same history effects as in the experimental time series. Neither of the two comparison series exhibited a reduction in traffic accidents coincident with the change in the drinking age. In this way, these data ruled out the threat of history using the estimate-and-subtract method of design elaboration. Critical to the present discussion is the fact that the experimental and comparison time series (i.e., the first and second estimates in the method of design elaboration) were derived from different groups of participants. The first estimate came from 18- to 20-year-olds, and there were two second estimates, with one coming from younger and the second from older participants.

A similar example (already described in Section 9.8) is provided by Guerin and MacKinnon (1985), who assessed the effects of a state law requiring that children under

3 years of age wear seat belts while riding in automobiles. Using a comparative ITS design, Guerin and MacKinnon examined yearly traffic fatalities separately for 0- to 3-year-olds and 4- to 7-year-olds. Only the data for the 0- to 3-year-olds should have shown an effect from the change in the law, while both time series should have been susceptible to the same history effects. The comparison time series provided no evidence of the effects of history in the form of a pseudo treatment effect. In this way, the effects of history were ruled out using the estimate-and-subtract method of design elaboration. Again, the first and second estimates of effects were from different participants.

11.3.2 Different Times

Instead of a second estimate derived from a separate group of participants, the second estimate used in the methods of design elaboration could be derived from a separate period of time. For example, Ross, Campbell, and Glass (1970) assessed the effects of a crackdown on driving under the influence of alcohol in Great Britain by comparing changes in traffic casualties before and after the crackdown began using a comparative ITS design. Two time series were created. The experimental time series (i.e., the first estimate) assessed casualties during weekend nights when the pubs were open, which is where most drinking and driving is presumed to occur. The comparison series (i.e., the second estimate) assessed casualties during the hours the pubs were closed. The experimental series should exhibit the effect of the crackdown plus any effects of history, and the comparison series should exhibit only the effects of shared history. There was a substantial interruption in the experimental series corresponding with the crackdown and no interruption in the comparison series at the same time. Hence, shared history effects were ruled out using the estimate-and-subtract method of design elaboration where the data in the second estimate were derived from a different time period than the time period from which the data in the first estimate were derived.

Another example (which was also described in Section 8.6.4) comes from Riecken et al. (1974) who reported a study to assess the effects of Medicaid on visits to the doctor (also see Cook & Campbell, 1979; Lohr, 1972). When Medicaid was first introduced, it paid for doctors' visits for families whose annual incomes were less than $3,000. A regression discontinuity design revealed a discontinuity corresponding to a cutoff score of $3,000, but other factors might have accounted for the discontinuity. As Cook and Campbell (1979) suggest, perhaps pregnant women or large families (with young children) are disproportionately represented among those with lower incomes, which could account for a greater number of doctor visits. To address these alternative explanations, the regression discontinuity design was repeated with data from the year before Medicaid was begun. These data showed no discontinuity at the cutoff score of $3,000, thereby suggesting that the discontinuity in the original data was due to Medicaid rather than to the alternative hypothesis. In this case, the data from the first and second estimates came from different time periods.

11.3.3 Different Settings

The methods of design elaboration can be implemented using data for a first and second estimate derived from different settings. For example, again consider the study by Reynolds and West (1987) which assessed the effects of a campaign to increase the sale of lottery tickets in convenience stores in Arizona. A time series of data from stores that implemented the campaign exhibited an increase in sales. A comparison series of data from stores that did not implement the campaign exhibited no change in sales, thereby ruling out shared threats to internal validity such as history effects. In this way, Estimate_1 and Estimate_2 were derived from different settings, namely, different convenience stores.

Anderson (1989) reports a study of the effects of a law mandating that seat belts be worn in the front seat of automobiles but not in the back seat or in other vehicles. The experimental data were derived from casualties due to traffic accidents of those in the front seats of automobiles. Comparison data were derived from casualties due to traffic accidents (1) of those in the back seats of automobiles and (2) of those in other types of vehicles besides automobiles. Both sets of series should be susceptible to similar history effects but only the experimental data should exhibit effects of the law. In this way, data from two different settings (front seats of automobiles versus back seats of automobiles and different types of vehicles) were used to implement the estimate-and-subtract method of design elaboration.

In multiple baseline designs, behavior modification studies often compare data from the same participant in different settings such as the home, school, and community where behavior incentives are implemented in one setting at a time to create a switching replication design (Kazdin, 2011). In such examples, data from different settings are used to implement the estimate-and-subtract method of design elaboration.

11.3.4 Different Outcome Measures

Just as the methods of design elaboration can use first and second estimates from different participants, times, and settings, the methods of design elaboration can also use first and second estimates from different outcome measures. In the Campbellian nomenclature, the outcome measure used in the second estimate is called a nonequivalent dependent variable (Shadish et al., 2002).

A nonequivalent group design by Braucht et al. (1995) which assessed the effects of treatment services for a group of homeless alcohol abusers provides an example. The program of services was aimed at reducing alcohol use. As expected, the analysis of these data revealed a reduction in alcohol use that was suggestive of a treatment effect, but the analysis could have suffered from hidden selection differences. For example, those who chose to participate in the intervention might have been more motivated, a selection difference that might not have been well accounted for in the data analysis. A measure of the quality of participants' relationships with their family was also collected

at both pretest and posttest, and a parallel analysis was conducted with these data. These data should not have exhibited much, if any, program effect because the treatment services did not address mending family relationships. But the data would likely be susceptible to similar selection differences due to differences in motivation. Null results were obtained with the alternative measure of relationships with family. So, to the extent that selection differences were the same in the two analyses, analyses of different outcome measures were used via the estimate-and-subtract method of design elaboration to disentangle the effects of bias from the effects of the treatment.

As previously mentioned, Reynolds and West (1987) assessed the effects of a campaign to increase sales of lottery tickets in convenience stores. In addition to collecting data on lottery sales, Reynolds and West also collected data on sales of other items at convenience stores (e.g., sales of groceries, cigarettes, and gasoline). The lottery ticket data were susceptible to influences due to both the treatment intervention and history. The alternative measures were susceptible only to the influences of history. In simple pretest–posttest comparisons, the sales of lottery tickets increased greatly, while the sales of the other items did not. In this way, different outcome measures were used as Estimate_1 and Estimate_2 in the estimate-and-subtract method of design elaboration.

McSweeny (1978) assessed the effects of charging a fee for telephone calls to directory assistance by collecting a time series of observations for the city of Cincinnati. (For those readers who are familiar only with cell phones, the McSweeny study was conducted before the widespread use of cell phones, which was a time when people could be charged for calls to telephone operators for assistance in finding telephone numbers or placing calls.) The fee Cincinnati imposed was for assistance only to find and place *local* calls. To rule out history threats to internal validity, McSweeny (1978) added a comparison time series for the use of directory assistance to find and place *long* distance calls (where no fee was imposed), which would be free of treatment effects but share the effects of history. So, again, the two estimates in the method of elaboration were derived from different outcome measures.

11.3.5 Multiple Different Size-of-Effect Factors

The preceding sections illustrate how the first and second estimates in the methods of design elaboration can be derived from different participants, times, settings, or outcome measures. In all these examples, a single first and second estimate were used based on differences in participants, times, settings, or outcome measures; but this need not be the case. Biglan et al. (2000) provides an example using multiple first and second estimates derived from different participants, different times, and different outcome measures. Biglan et al. (2000) were interested in the effects of raising the minimum drinking age that took place in various states during the 1970s and 1980s. The study collected time-series data on traffic accidents and disaggregated the data on multiple dimensions, including age (different participants), time of day (different times), and type of crash (different outcome measures). Alcohol is more often a factor in drinking-related

traffic accidents in certain situations than in others. For example, alcohol is more often a factor in single-vehicle nighttime accidents and less often a factor in multiple-vehicle daytime accidents. As a result, Biglan et al. (2000) expected to find differential effects of changes in the minimum drinking ages across the different disaggregated time series. Biglan et al. (2000), however, expected most history effects to be the same across the different comparison conditions. In this way, multiple comparison series from different sources were used to rule out the threat to internal validity due to history effects using the estimate-and-subtract method of design elaboration.

Another example comes from a study by Khuder et al. (2007). The study assessed the effect of a ban on smoking in a city in Ohio. Data were collected on hospital admissions for coronary heart disease (CHD) over a 6-year period, with the ban being introduced in Year 3. The data exhibited a decrease in admissions for CHD. To control for threats to internal validity due to history effects using the estimate-and-subtract method of design elaboration, two types of comparison time series were added. One comparison series consisted of data from a different setting—hospitals in another city that imposed no smoking ban. The other comparison series consisted of data from the same hospital on the nonequivalent dependent variable of hospital admissions for nonsmoking related illnesses. Neither comparison series exhibited declines at the same time the smoking ban was introduced, hence adding credence to the conclusion that the ban reduced CHD admissions.

11.4 CONCLUSIONS

The methods of design elaboration provide a means of taking account of threats to internal validity and the biases they produce. The methods of design elaboration remove biases in an estimate of a treatment effect by adding a second estimate to a first one. Together, the two estimates can disentangle the treatment effect from the bias. I have described three ways to combine the two estimates called the three methods of design elaboration. When an estimate is added to rule out a given threat to validity, it operates according to one of these three methods of elaboration. To implement the three methods, a researcher thinks through ways in which a second estimate can be added to a first estimate, so that the two estimates have different amounts of either the treatment effect or the bias. The most commonly used method of design elaboration is the estimate-and-subtract method. Considering only that one method, however, needlessly limits the possible design options and the possibility of crafting the best design for the given circumstances.

In the methods of design elaboration, the two estimates used to disentangle the treatment effect from bias must each come from a different source where the sources are defined by the four size-of-effect factors. That is, the different sources for the two estimates in design elaboration are different participants, times, settings, or outcome measures. Arguably the most common source for a second estimate is a different group

of participants. But to consider only adding an estimate from a different group of participants is, again, to needlessly limit the possible design options and the possibility of crafting the best design for the given circumstances. Researchers should therefore consider all possible options, including options from the principle of parallelism when implementing the methods of design elaboration. When thinking of how to add a second estimate to disentangle a bias from a treatment effect, consider how the second estimate might be derived from different participants, times, settings, or outcome measures.

11.5 SUGGESTED READING

Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Vol. 2, pp. 89–115). Thousand Oaks, CA: SAGE.

Reichardt, C. S., & Gollob, H. F. (1989). Ruling out threats to validity. *Evaluation Review, 13,* 3–17.
—Provide a more complete explication of methods of design elaboration.

Shadish, W. R., & Cook, T. D. (1999). Comment—Design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science, 14*(3), 294–300.
—Provides an extensive list of design features that can be used to rule out threats to validity by the methods of design elaboration.

12

Unfocused Design Elaboration and Pattern Matching

The original hypothesis and the alternative artifactual explanation associated with each threat to internal validity are expected to produce different patterns of results; the obtained results are compared to each of the expected patterns to determine which provides the best account of the data—a process termed *pattern matching*.

—West, Cham, and Liu (2014, pp. 61–62; emphasis in original)

The more numerous and independent the ways in which the experimental effect is demonstrated, the less numerous and less plausible any singular rival invalidating hypothesis becomes.

—Campbell and Stanley (1963, p. 206)

The fundamental postulate of multiplism is that when it is not clear which of several options for question generation or method choice is "correct," all of them should be selected so as to "triangulate" on the most useful or the most likely to be true. . . . When results are stable across multiple potential threats to causal inference, internal validity is enhanced.

—Cook (1985, pp. 38, 46)

When the principal source of ambiguity comes from unmeasured biases, a valuable replication tries to study the same treatment effect in the presence of different potential sources of unmeasured biases. If repeatedly varying the most plausible sources of bias does little to alter the ostensible effects of the treatment, then the evidence in favor of an actual treatment effect is gradually strengthened. . . .

—Rosenbaum (2015b, p. 26)

Overview

The difference between focused and unfocused design elaboration is explicated. Chapter 11 described focused design elaboration, which uses multiple estimates to address a shared threat to internal validity. In contrast, unfocused design elaboration uses multiple estimates to address separate threats to internal validity. The process of pattern matching accounts for the functioning of both focused and unfocused design elaboration.

12.1 INTRODUCTION

The present chapter distinguishes between focused and unfocused design elaboration. The previous chapter was concerned with focused design elaboration. By **focused design elaboration,** I mean design elaboration whereby two estimates are used to disentangle the treatment effect from the effect of a shared threat to internal validity. That is, in focused design elaboration, both of the two estimates (the first and the second estimates) are influenced by the same threat to internal validity. For example, in the estimate-and-subtract method of design elaboration, the first estimate is equal to a treatment effect plus the effect of a threat to validity, and the second estimate is a measure of the effect of the same threat to validity (though not of the effect of the treatment). More than two second estimates could be used in a given focused design elaboration, but they would all be influenced by the same threat to validity.

Unfocused design elaboration is different. In unfocused design elaboration, a second estimate of the treatment effect is added to a first estimate of the treatment effect, but the second estimate can be subject to one or more separate threats to internal validity. That is, each estimate of the treatment effect can be subject to different threats to internal validity. In addition, more than two estimates of a treatment effect (i.e., estimates besides the first and second estimate) could be combined in unfocused elaboration, but each estimate would be subject to different threats to internal validity. To the extent that the results from all the estimates agree, it becomes increasingly less likely, as the number of estimates increases, that the results are all due to different threats to internal validity rather than that the estimates all represent the effect of the treatment. So taken together, the multiple estimates can provide stronger evidence for a treatment effect than any one estimate by itself (Rosenbaum, 2015a, 2015b, 2017). According to Rosenbaum's (2015b, 2017) nomenclature, when derived from a single study, the multiple estimates in unfocused design elaboration are called evidence factors.

To repeat: In focused design elaboration, both the first and the second estimates are subject to the same threat to internal validity. In contrast, in unfocused design elaboration, each estimate of a treatment effect is subject to different threats to internal validity. The greater the number of estimates and the more they agree, the less likely it is that the series of different threats would be operating in the same pattern and the more likely it is that the results represent a shared treatment effect. Consider four examples of unfocused design elaboration.

12.2 FOUR EXAMPLES OF UNFOCUSED DESIGN ELABORATION

The first example comes from a study by Lipsey et al. (1981) that estimated the effects of a juvenile diversion program using five different research designs, each of which was susceptible to its own, independent threats to validity. Before the diversion program was implemented, juveniles who were arrested for a crime were either counseled and

released or sent to the probation department within the juvenile court system. The juvenile diversion program introduced a third option where juveniles arrested for crimes were referred to social services. The diversion program was intended mainly as an alternative to probation but could also be used as an alternative to the counsel-and-release disposition. Lipsey et al.'s (1981) first design was a nonequivalent group comparison involving juveniles at only some of the sites in the study. In this design, juveniles who entered the diversion program and received a full complement of social services were compared to juveniles at the same sites who entered the diversion program but did not receive the full complement of services. The results showed that those who received the full complement of services had lower recidivism rates than those who did not receive the full complement of services.

Lipsey et al.'s (1981) second design was a tie-breaking randomized experiment where two sites in the study agreed to assign certain categories of juveniles at random to the diversion program, the counsel-and-release disposition, or probation in the court system. Differences in recidivism did not reach statistical significance but favored the juveniles sent to the diversion program compared to those given a counsel-and-release disposition.

Lipsey et al.'s (1981) third design implemented a regression discontinuity design at two other sheriff's stations, each using two cutoff scores to create three treatment groups. The QAV was a disposition scale composed of 11 separate ratings by police officers of the need to assign juveniles to probation. Those high on the QAV were assigned to probation using a high cutoff score; those low on the QAV were assigned to the counsel-and-release disposition using a low cutoff score; and those with scores in between the two cutoff scores were assigned to the diversion program. Discontinuities in the regression of recidivism on the QAV at both cutoff scores favored the diversion program (though only one discontinuity was statistically significant).

The fourth design was a nonequivalent group comparison using matched samples of juveniles assigned to different treatments, but the sample sizes were too small for differences in recidivism to reach statistical significance.

The fifth estimate of a treatment effect was derived from an interrupted time-series design that assessed arrest records before and after the diversion program was implemented. The results revealed that the diversion program was associated with a decreasing trend in arrests.

Note that each of the five designs was susceptible to substantial uncertainty due to threats to internal validity, but the threats to internal validity in each of the treatment effect estimates were different. The results from the five studies could have been due to the effects of independent threats to validity or to the same shared effects of the treatment. That so many results were in substantial agreement strengthened the conclusion that the shared treatment was responsible, rather than the independent threats. So taken as a whole, the set of estimates provided stronger evidence for the presence of a treatment effect than any one estimate alone. The multiple estimates of the treatment effect worked together to strengthen each other. This is unfocused design elaboration.

The second example of unfocused design elaboration comes from a study by Reynolds and West (1987), which has already been described (see Sections 6.4, 11.3.3, and 11.3.4). As previously noted, the study assessed the effects of a campaign to increase the sales of lottery tickets at convenience stores in Arizona. Reynolds and West (1987) estimated the effect of the campaign using four designs. With a nonequivalent group design, they compared 44 stores that implemented the sales campaign with 44 matched stores that did not implement the sales program. The stores that implemented the campaign subsequently sold substantially more tickets than the stores that did not implement the campaign, even though both sets of stores sold the same amount of tickets on a pretest measure. In the second design, Reynolds and West (1987) used a pretest–posttest comparison with a nonequivalent dependent variable and found that lottery ticket sales increased in the stores that implemented the campaign, while sales of other products (e.g., groceries, cigarettes, and gasoline) remained the same at those same stores. A third study used a short interrupted time series (ITS) design with a comparison time series and found an increase in sales of lottery tickets corresponding to the start of the campaign in stores that implemented the campaign but no change in sales in stores that did not implement the campaign. Finally, Reynolds and West (1987) used an ITS design with removed treatment interventions to compare sales in groups of stores that started and then stopped the sales campaign. The four designs were susceptible to different threats to internal validity, and yet they all produced estimates of a substantial treatment effect. That all four designs led to the same conclusion about the effect of the sales campaign strengthened that conclusion.

The third example comes from Levitt and Dubner (2005). This study argued that the 1973 *Roe v. Wade* Supreme Court decision, which legalized abortion, caused a substantial drop in crime rates throughout the United States beginning in the early 1990s. Note that crime had risen from the 1970s to the early 1990s, after which it began to decline. Levitt and Dubner (2005) hypothesized that, if abortion was causing a decline in the crime rate, the decline should follow after a 20-year lag. This is because abortion would have reduced the number of unwanted children whose unhappy upbringing might have caused them to pursue criminal activity. The pool of active criminals would not show a decline for about 20 years—not until the "missing" children would have become young adults had they been born. Levitt and Dubner used several pieces of evidence to support their conclusion.

1. Some states legalized abortion under certain circumstances (such as rape and incest) in the late 1960s. These states exhibited a drop in crime earlier than the states in which abortion was legalized only in 1973 following the *Roe v. Wade* decision.
2. There was a negative correlation across states between the rate at which abortions took place and the rate of crime 20 years later. But there was little correlation in time periods less than 20 years following *Roe v. Wade*.
3. The decline in the crime rate in the 1990s was due mostly to crimes that are most likely to be committed by younger rather than older adults.

4. The same relationships between increases in abortion and decreases in crime rates were found in Australia and Canada.
5. After abortions were banned or made harder to obtain in Europe, there was an increase in crime 20 years later.

None of these results, taken alone, is definitive—each is susceptible to alternative interpretation. But having so many results that support the conclusion that abortion lowers the crime rate adds to that conclusion's credibility.

The fourth example comes from Yin's (2008) report on the evaluation of partnerships between colleges and high schools for teaching science in the high schools. In that study, the performances of 18 high schools were compared in four science areas or strands. In the high school labeled as SHAR, a partnership had "helped the classrooms to strengthen their science instruction" in strands 1 and 3 (but not in strands 2 and 4). In the high school labeled as LNES, a partnership had strengthened the curriculum in strands 2 and 4 (but not in strands 1 and 3). None of the other schools had participated in a partnership. School SHAR performed at the average of all the other 16 schools on strands 2 and 4, but on strands 1 and 3, school SHAR performed way beyond all the other schools. Complementary results were obtained for school LNES—that school performed at the average of the other schools on strands 1 and 3 but performed way beyond the other schools on strands 2 and 4. Note that the comparison among the schools is a nonequivalent group comparison and the comparison of different strands is a comparison of nonequivalent dependent variables. Either comparison by itself would not produce compelling evidence for a treatment effect. But the complex pattern of results produced when the two comparisons are combined is difficult to explain in any way other than that it reveals a positive effect of the partnership program.

12.3 PATTERN MATCHING

In the most general terms, estimating the effect of a treatment effect involves matching patterns. To estimate a treatment effect, a researcher creates a pattern of results so that the effect of the treatment can be distinguished from the effects of threats to internal validity (Abelson, 1995; Cook & Campbell, 1979; Scriven, 1976; Shadish et al., 2002; Trochim, 1985, 1989; West et al., 2014). That is, the researcher collects data wherein the treatment is predicted to result in certain patterns of outcomes (should a treatment effect be present), while threats to internal validity are predicted to result in alternative patterns of outcomes (should they be present). The researcher then compares the predicted patterns to the data that are obtained. To the extent that the pattern predicted by the treatment fits the data better than do the patterns predicted by threats to internal validity, the treatment is declared the winner and a treatment effect is plausible. This is the procedure of pattern matching, which is the mechanism by which both focused and unfocused design elaboration operate.

Often, the best patterns for distinguishing treatment effects from the effects of threats to internal validity are complex. More complex patterns allow for more differentiated predictions between treatment effects and the effects of threats to internal validity. For example, fingerprints and DNA evidence are so telling in criminal investigations because both fingerprints and DNA can provide sufficiently complex patterns that they rule out all but one person as the perpetrator of a crime. In quasi-experimentation, complex patterns are obtained through both focused and unfocused design elaboration in the form of multiple estimates of treatment effects.

The value of complex (or elaborate) patterns has long been recognized. For example, Cochran (1965, p. 252; 1972) described Fisher's explication of the role of elaborate patterns in research design thusly:

> When asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate." The reply puzzled me at first, since by Occam's razor the advice usually given is to make theories as simple as is consistent with the known data. What Sir Ronald meant, as the subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many *different* consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold. (emphasis in original)

Similarly, Cook and Campbell (1979, p. 22) explained that plausible alternative hypotheses can be ruled out "by expanding as much as we can the number, range, and precision of confirmed predictions. The larger and more precise the set, the fewer will be the alternative explanations. . . ." Shadish et al. (2002, p. 105) expressed the same idea: "The more complex the pattern that is successfully predicted, the less likely it is that alternative explanations could generate the same pattern, and so the more likely it is that the treatment had a real effect." The purpose of both focused and unfocused design elaboration is to produce a complex pattern of results that can be plausibly explained only as the result of a treatment effect.

Like focused design elaborations, unfocused design elaborations can be implemented in accordance with the principle of parallelism. That is, the multiple treatment effect estimates in unfocused design elaboration can be obtained by comparing different participants (e.g., multiple comparison groups), different times (e.g., ITS comparisons), different settings (e.g., treatments implemented in different locations), or different outcome measures (e.g., nonequivalent dependent variables).

12.4 CONCLUSIONS

As Fisher suggested, researchers should think through all the ways in which a treatment can have effects and try to study as many of these ways as possible to produce a complex pattern of results that cannot be plausibly explained by threats to internal validity. A complex pattern of results can be obtained by either focused or unfocused design elaboration. In focused design elaboration, the researcher thinks of ways in which the

treatment and a specified threat to internal validity can be placed in competition using two separate estimates where either the treatment effect varies across the estimates or the effect of the shared threat to internal validity varies across the estimates. In creating such comparisons, the researcher keeps in mind that the two estimates can be obtained from different participants, times, settings, or outcome measures. In unfocused design elaboration, the researcher thinks of different ways in which the treatment can be estimated where each way is susceptible to different threats to internal validity. To the extent that different threats to validity are operating across comparisons (i.e., to the extent that there is a "heterogeneity of irrelevances") and yet the results of the comparisons converge on the same estimate of the treatment effect, that estimate becomes more plausible (Cook, 1990). In creating such complex comparisons, the researcher keeps in mind that the multiple treatment effect estimates can be obtained by contrasting different participants, times, settings, or outcome measures. Indeed, the multiple estimates can be derived from any of the comparisons listed in Table 10.1.

12.5 SUGGESTED READING

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 81–106). New York: Holt, Rinehart & Winston.
—Explains how pattern matching is central to all knowledge acquisition.

Cook, T. D. (1985). Post-positivist critical multiplism. In R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21–62). Beverly Hills, CA: SAGE.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. In L. B. Sechrest & A. J. Figueredo (Eds.), *Program evaluation: A pluralistic enterprise* (New Directions in Program Evaluation No. 60, pp. 13–57). San Francisco: Jossey-Bass.
—Provide an overview of the logic and practice of critical multiplism, which is a strategy that includes both focused and unfocused design elaboration.

Lipsey, M. W., Cordray, D. S., & Berger, D. E. (1981). Evaluation of a juvenile diversion program: Using multiple lines of evidence. *Evaluation Review, 5,* 283–306.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11,* 691–714.
—Provide excellent empirical examples of unfocused design elaboration.

Rosenbaum, P. R. (2015a). Cochran's causal crossword. *Observational Studies, 1,* 205–211.

Rosenbaum, P. R. (2015b). How to see more in observational studies: Some new quasi-experimental devices. *Annual Reivew of Statistics and Its Applications, 2,* 21–48.

Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference.* Cambridge, MA: Harvard University Press.

Trochim, W. M. K. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9,* 575–604.
—Explain the logic and practice of pattern matching.

13

Principles of Design and Analysis for Estimating Effects

You can't fix by analysis what you bungled by design.
—LIGHT, SINGER, AND WILLETT (1990, p. viii)

We should draw causal inferences where they seem appropriate but also provide the reader with the best and most honest estimate of the uncertainty of that inference.
—KING, KEOHANE, AND VERBA (1994, p. 76)

A compelling observational study is one that has received the implicit endorsement of surviving critical discussion largely unscathed.
—ROSENBAUM (2017, p. 217)

Intervention research is difficult.
—SHERIDAN (2014, p. 299)

Overview

The present chapter describes 14 principles to follow when crafting research designs and performing statistical analyses to estimate treatment effects.

13.1 INTRODUCTION

Occasionally, you will come across statements that randomized experiments are required if you are to estimate treatment effects. For example, Ashenfelter and Card (1985, p. 648) wrote, "we conclude that randomized clinical trials are necessary to determine program effects." I hope I have disabused you of such a belief. Randomized experiments can certainly be useful under the proper conditions, but they are not required to draw causal inferences. This volume explains the logic by which treatment effects can be estimated. Nowhere does that logic specify that randomized experiments

are the only acceptable method. The logic of estimating effects can be implemented with quasi-experiments as well as with randomized experiments.

Similarly, you will occasionally come across statements that appear to claim that quasi-experiments are never better than randomized experiments. For example, Cook and Wong (2008a, p. 159) stated, "Because of the randomized experiment's more elegant rationale and transparency of assumptions, no quasi-experiment provides a better warrant for causal inference." And Cook (2003, p. 114) wrote, "even if [randomized experiments] are not perfect in research practice, this article shows how they are logically and empirically superior to all currently known alternatives."

I hope I have also disabused you of such beliefs, if they are taken to mean that randomized experiments are always superior to other methods. Randomized experiments are, indeed, often superior to quasi-experiments. Even when they suffer from difficulties such as noncompliance to treatment assignment and differential attrition from treatment conditions, which can undermine their credibility, randomized experiments are still often superior to even the best quasi-experiments. This is not always the case, however. Sometimes quasi-experiments can be implemented with greater fidelity than randomized experiments and can lead to more credible causal inferences.

At the same time, just because randomized experiments are not required nor always superior to quasi-experiments does not mean that anything goes—that any approach to estimating effects will be just as good as any other. In specific research settings, some methods will be better than others. Unfortunately, there is no cookbook that will guarantee that the best research methods can be chosen for a given research setting. The best methods for estimating effects can, however, be crafted by following certain principles. What follows is an explication of 14 fundamental principles for devising the most credible designs and analyses for estimating effects. These principles apply to both randomized and quasi-experimental designs.

13.2 DESIGN TRUMPS STATISTICS

Shadish and Cook (1999, p. 300) wrote: "When it comes to causal inference from quasi-experiments, design rules, not statistics." These authors mean that research design, rather than statistical analyses, should be the first line of defense against threats to internal validity. Implementing a hastily chosen research design and attempting to correct for its inadequacies with fancy statistics are likely to be less successful strategies than anticipating and controlling for threats to internal validity by using carefully-thought-out design features (Cook & Wong, 2008a; Shadish et al., 2002; Rosenbaum, 1999, 2017). Choose a design to avoid threats to validity rather than allowing threats to operate freely and trying to remove them statistically. Choose a design so that the threats to validity that cannot be avoided can be addressed using simple (though complexly patterned) comparisons rather than elaborate statistical machinations. Choose

a design where the assumptions of statistical procedures are credible rather than dubious. The point is this: when estimating treatment effects, choose a design thoughtfully to minimize reliance on uncertain statistical analysis for addressing threats to internal validity.

13.3 CUSTOMIZED DESIGNS

The constraints placed on a research design differ greatly across research settings. Different constraints and circumstances demand different research designs, and the researcher must fit the research design to the specifics of the research setting. Unfortunately, discussions of quasi-experiments often focus on only a limited number of prototypical research designs. As a result, researchers often conceptualize their task as that of choosing among a limited set of prespecified designs. This is a mistake. Researchers should not choose from among a fixed collection of prefabricated designs, but should rather craft a customized design by combining the best features from a large pool of design options. Only by using tailor-made designs can researchers hope to best accommodate the demands and constraints of the specific research setting. Do not take a prototypical design off the shelf. Instead, carefully measure the research circumstances and cut the design cloth to fit.

Just as the expert tailoring of a suit requires multiple measurements and remeasurements, the process of expert research design selection should proceed iteratively. Tentatively choose a basic comparison and assess its strengths and weaknesses. Then consider how the basic comparison might be supplemented or elaborated to capitalize on strengths and minimize weaknesses. Next consider further or alternative supplements and elaborations. Then start with a different basic comparison and iteratively add supplements and elaborations to that design to see where that takes you. And so on. From among all the different options, choose the design (or designs) that will likely produce the most credible conclusions given the constraints of the research setting. Such a process of design creation is time consuming—certainly more so than simply selecting from a list of prespecified, prototypical designs. But the time spent will return dividends in the precision and credibility of the conclusions reached. The point is: don't implement a design until you have carefully considered the full range of possible designs and design features (Rosenbaum, 2015b, 2017; Shadish & Cook, 1999).

13.4 THREATS TO VALIDITY

A threat to validity is an alternative explanation for the results of a study. According to Shadish et al. (2002), there are four classes of threats to validity: construct, internal, statistical conclusion, and external validity. *Construct validity* has to do with whether the

size-of-effect factors in a study are correctly labeled. If one or more of the size-of-effect factors is mislabeled, the conclusion suffers from construct invalidity. *Internal validity,* a special case of construct validity, concerns the labeling (or mislabeling) of only the size-of-effect factor of the treatment or cause. It also concerns only mislabelings of the treatment (or cause) that are confounds that could have been present even if the treatment had not been implemented. Other confounds (such as placebo effects) that are a result of the treatment implementation are threats to construct validity. *Statistical conclusion validity* addresses two questions: (1) Is the degree of uncertainty that exists in the estimate correctly represented? and (2) Is that degree of uncertainty sufficiently small (e.g., is the estimate of the treatment effect sufficiently precise, and is the power of statistical significance tests sufficiently great?)? *External validity* has to do with the intended purpose of the study and with the generalizability of the study results. External validity asks whether the researcher studied the size-of-effect factors of interest (the cause, participants, times, settings, and outcome measures) and, if not, whether the results that were obtained generalize to the size-of-effect factors of interest.

Researchers must take account of threats to validity if they are to draw confident and credible conclusions about the effects of treatments. Researchers should carefully assess the threats to validity that are likely to be present in each design being considered. Then researchers should try to choose a design or add features to a design to render the identified threats implausible.

In this volume, I have focused on internal validity and statistical conclusion validity because threats to internal and statistical conclusion validity are what most distinguish different quasi-experimental designs. Threats to internal and statistical conclusion validity, and their means of control, differ significantly across quasi-experimental designs. In contrast, threats to construct and external validity, and the means to address them, are relatively more similar across designs. Nevertheless, researchers need to be concerned with all four types of threats to validity (described in more detail in Chapter 3).

Different designs suffer differentially from threats to the four types of validity. Which threats to validity are most plausible depends not just on the research design but also on how the research design interacts with the specifics of the research setting. A randomized experiment that has high internal validity in one setting might, because of noncompliance to treatment assignment or differential attrition from treatment conditions, have low internal validity in another. The strengths of a design in one setting might be weaknesses in another. Researchers should craft research designs to fit the threats to validity in the specific research setting.

In addition, there are usually trade-offs among the four types of validity. For example, randomized experiments often enjoy superior internal validity but can suffer from external invalidity because of the limited settings in which randomized can be implemented. For example, if only certain types of institutions (e.g., schools, businesses, hospitals) would allow a randomized experiment to be implemented, then the results may not generalize to the types of institutions that would not allow randomized

experiments. Conversely, nonequivalent group designs can often be implemented in a broader range of situations than can randomized experiments. But nonequivalent group designs generally suffer more from threats to internal validity than do randomized experiments. It is usually impossible to maximize all four types of validity in a research design. Instead, researchers will have to prioritize the types of validity and seek to craft designs that maximize the types of validity at the top of the list. To address all sources of invalidity, we will have to rely on results from multiple studies, which even then will never be perfectly free from alternative interpretation.

13.5 THE PRINCIPLE OF PARALLELISM

As described in Section 10.2, the principle of parallelism states that if a design option is available for one of the four size-of-effect factors of participants, times, settings, and outcome measures, a parallel design option is available for the other three size-of-effect factors. For example, if a comparison of different participants could be added to a design, a comparison could (at least in theory) alternatively or additionally be added that is comprised of different times, settings, or outcome measures.

Attending to the principle of parallelism can help researchers extend their reach beyond the most obvious design options. For example, it is often obvious that a design can be based on (or elaborated by adding) a comparison of different participants. It might not be as obvious that a design can be based on (or elaborated by adding) a comparison of different times, settings, or outcome measures. The principle of parallelism helps researchers recognize all the types of design features that are possible.

13.6 THE TYPOLOGY OF SIMPLE COMPARISONS

As described in Sections 10.3–10.6, the effect of a treatment can be estimated using one of four basic comparisons. These four comparisons are derived by varying the treatment across participants, times, settings, or outcome measures. In each of these comparisons, treatments can be assigned in one of three ways: random assignment, nonrandom assignment based on an explicit quantitative assignment rule, or nonrandom assignment not based on a quantitative assignment rule. Crossing the four types of comparisons with the three types of treatment assignments produces a typology of 12 simple comparisons (see Table 10.1). This typology is a consequence of the principle of parallelism.

Researchers should consider all 12 of the simple comparisons (and combinations of these comparisons) when designing a study. Some of the comparisons will likely not be possible in the given research settings, while some of those that are possible might be overlooked by researchers who do not explicitly consider all 12 possibilities.

13.7 PATTERN MATCHING AND DESIGN ELABORATIONS

Estimating treatment effects is an exercise in pattern matching. To estimate a treatment effect, a researcher must create a pattern of results wherein the effect of the treatment can be distinguished from the effects of threats to internal validity. That is, the researcher must collect data where the treatment is expected (should it have an effect) to produce a pattern of results that differs from the pattern of results expected to be produced by threats to internal validity. The researcher then compares the different patterns to the obtained results to see which explanation provides the best match.

Often, the best patterns for distinguishing treatment effects from the effects of threats to internal validity are complex patterns—much as the complex patterns of fingerprints and DNA help discriminate among criminal suspects. Patterns of results can be made complex by both focused and unfocused design elaborations.

There are three types of focused design elaboration: the estimate-and-subtract method, the vary-the-size-of-the-treatment-effect method, and the vary-the-size-of-the-bias method. In these three methods, two estimates (a first and a second estimate) are used to disentangle a treatment effect from the effects of a specified threat to internal validity. Both the first and the second estimates are influenced by the same identified threat to internal validity, but the second estimate is influenced by a different combination of effects than the first estimate. In the estimate-and-subtract method of design elaboration, for example, the second estimate is a separate estimate of a threat to internal validity that is identified as plausibly present in the first estimate. In accord with the principle of parallelism, each of these three methods of focused design elaboration can be implemented using, in the two estimates, different participants, times, settings, or outcome measures. A vast array of potential designs can be fashioned by combining the different types of focused design elaborations with each of the 12 simple comparisons in Table 10.1.

In addition to focused design elaboration, there is also unfocused design elaboration. In unfocused design elaboration, multiple estimates of the treatment effect are obtained that are susceptible to different threats to internal validity. To the extent that the results from all the estimates of the treatment effect agree, it becomes increasingly less likely, as the number of independent estimates increases, that the results are all due to independent threats to validity than that they represent the effect of the treatment. So taken together, the multiple estimates can provide stronger evidence for a treatment effect than any one estimate by itself. The types of estimates that can be combined include each of the 12 simple comparisons in Table 10.1.

In designing quasi-experiments, researchers should try to prevent threats to internal validity from arising and use design elaborations to address those threats to validity that do arise. That is, researchers should try to take account of identified threats to internal validity using focused design elaboration. Researchers should also use unfocused design elaboration to cope with threats to internal validity that cannot be addressed with focused design elaboration.

Many of the design options described in this volume are not widely appreciated in the literatures in either statistics or the social sciences. My hope is that the typology of simple comparisons and the discussion of both focused and unfocused design elaboration will help readers appreciate a wider range of design options than they would otherwise. Having a wider range of design options available can only lead to better research designs.

13.8 SIZE OF EFFECTS

Whether an effect is zero or nonzero is the question of the "existence" of the effect. Whether an effect is positive or negative is the question of the "direction" of the effect. The numerical magnitude of an effect is a question of the "size" of the effect. Knowing the direction of an effect reveals the existence of the effect. So, knowing the direction of an effect is more informative than knowing only its existence, and knowing the size of an effect reveals both existence and direction. Thus, knowing the size of an effect is more informative than knowing only the existence or direction of the effect.

Even though size is more informative than either existence or direction, many methodological writings focus on the existence or direction of an effect as much as, if not more than, on its size. For example, Campbell and Stanley (1966), Cook and Campbell (1979), and Shadish et al. (2002) often focus on existence and direction rather than on size (Reichardt, 2011b; Reichardt & Gollob, 1997). Evidence of this focus can be found, for instance, in the way Campbell and Stanley (1966, p. 5) define internal validity, which according to them asks the question of "Did in fact the experimental treatment make a difference in this specific experiment?" This is a question about the existence of an effect. A focus on size, instead, would mean that internal validity would ask "How much of a difference did the experimental treatment make in this specific experiment?"

In addition, a focus on size is evidenced not just in methodological writings. Research practices also often focus on existence and direction more than on size. For example, statistical significance tests focus on the existence and direction of an effect. That is, by itself, statistical significant tests reveal only that an effect is not equal to zero (e.g., whether it exists) or is either less than or greater than zero (e.g., its direction). Confidence intervals are necessary if the researcher wishes to appreciate the likely size of a treatment effect. The differences between the results of statistical significance tests and confidence intervals can perhaps best be appreciated with the use of pictures.

Figure 13.1 presents four lines on which are marked the results of 95% confidence intervals for the size of the treatment effect from four different studies (Reichardt & Gollob, 1997). The ends of the confidence intervals are marked with parentheses, and the zero point on each line is marked with a "0." In both lines A and B, the confidence intervals lie completely above the zero point. Hence, a statistical significance test using an alpha level of .05 would reject the null hypothesis that the effect size is equal to zero.

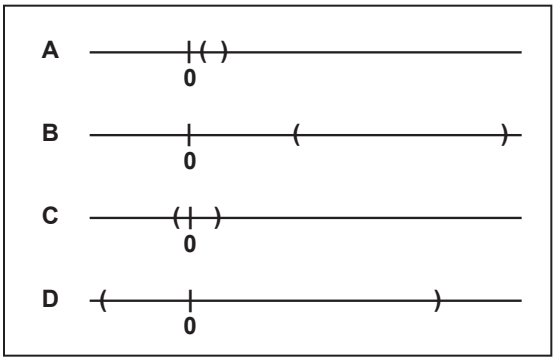


FIGURE 13.1. Confidence intervals for four possible outcomes.

In addition, the two statistical significance tests would have the same *p*-value and lead to the same degree of confidence about the direction of the treatment effect. (That the *p*-values would be the same is revealed by the fact that the midpoint of each confidence interval is four standard errors above the zero point.) But the confidence intervals for lines A and B reveal quite different things about the likely size of the treatment effects. In line A, the confidence interval reveals that the estimate of the treatment effect is relatively close to zero, while the confidence interval in line B reveals that the estimate of the treatment effect is relatively far from zero. The point is that these differences in size are clearly revealed by the confidence intervals but not by the *p*-values of the statistical significance tests (Wasserstein & Lazar, 2016).

Now consider the results portrayed by lines C and D in Figure 13.1. In these two instances, the 95% confidence intervals contain the value of 0, so results from statistical significance tests at the alpha level of .05 would not reject the null hypothesis that the effect size is equal to 0. In addition, the two statistical significance tests would have the same *p*-value because the midpoint of each confidence interval is one standard error above the zero point. But the confidence intervals for lines C and D reveal quite different things about the two sets of results. In line C, the confidence interval reveals that the estimate of the treatment effect is relatively close to zero while the confidence interval in line D reveals that the estimate of the treatment effect could well be relatively far from zero. These differences in size are clearly revealed by the confidence intervals but not by the *p*-values of the statistical significance tests.

Note that by convention, the results depicted by lines A and B could both be accepted for publication because they are both statistically significant at the .05 level. In contrast, the results depicted in lines C and D would tend not to be accepted for publication because they are both statistically nonsignificant. Admittedly, the results depicted in line D are not very informative. The effect size could be either negative or positive, and it could be quite small or very large. The results in line C are different, however. Though nonsignificant statistically, the results depicted in line C show that the effect is likely to be very small. In many cases, such results would be very informative and well

worth publishing. But it would take confidence intervals, and not just *p*-values, to be able to most clearly distinguish results in line C from results in line D.

Despite the relative advantages of confidence intervals, it is common research practice to emphasize statistical significance tests more than confidence intervals. This is a mistake. Researchers should focus on size of treatment effects and not just on existence or direction (Abelson, 1995; Reichardt & Gollob, 1997; Wilkinson & APA Task Force, 1999).

In contrast to this advice, some researchers have argued that size should not be the focus of attention because size can vary so greatly with the causes, participants, times, settings, and outcome measures. But to throw out information about size is a mistake precisely because we need to know how size varies with causes, participants, times, settings, and outcome measures. Size is important, even in tests of theories that predict no more than direction, because a small effect size is generally more plausibly interpreted as due to a threat to internal validity than is a large effect size. So research with small effect sizes should be interpreted more cautiously, even when testing theories only about the direction of effects, than research with large effect sizes. When attempting to ameliorate social problems, knowing how the size of an effect varies with causes, participants, times, settings, and outcome measures is critical for determining which treatments best fit which participants, times, settings, and outcome measures.

In addition to paying attention to the size of effects, researchers should also focus on the precision of treatment effect estimates. Not only do we seek unbiased estimates of effects, but we also need precise estimates of effects. Paying attention to both the size and precision of treatment effect estimates means, once again, paying attention to confidence intervals rather than to the results of statistical significance tests alone. One great benefit of focusing on confidence intervals more than on the results of statistical significance tests is that confidence intervals appropriately emphasize the uncertainty in our findings, where uncertainty is represented by the width of the confidence interval. Part of honestly reporting results is a forthright assessment of uncertainty. Although the results in lines A and B (see Figure 13.1) have the same *p*-values, they reveal very different things about the uncertainty with which the treatment effect is estimated. The same holds for lines C and D.

In addition to using confidence intervals, it is recommended that effect sizes be reported in terms that are substantively meaningful (Coalition for Evidence-Based Policy, 2003). As the Coalition (2003, p. 9) emphasizes, "standardized effect sizes may not accurately convey the educational importance of an intervention, and, when used, should preferably be translated into understandable, real-world terms. . . ." For example, rather than using traditional standardized effect sizes (such as Cohen's *d*), researchers could report, say, that an effect raised the average level of performance in the treatment group to between the 70th and 80th percentile of the comparison group. Similarly, Chambless and Hollon (2012) emphasize that standardized measures of effect sizes (such as Cohen's *d*) can be misleading because they are influenced by the reliability of the measures. The greater the reliability, the larger are the standardized effects, even if

the true size of the effect remains the same. Instead of standardized measures, Chambless and Hollon (2012, p. 537) suggest, in the context of clinical research, "reporting what percentage of patients in each group no longer meet diagnostic criteria for the primary diagnosis at the end of treatment or reporting what percentage of patients in each group meet criteria for clinically significant change." Chambless and Holland cite Jacobson and Truax (1991) at this point.

Researchers must also be careful not to present sanitized versions of their findings (Moskowitz, 1993). For example, researchers should not suppress negative findings and report only positive effects. To avoid reporting results that cannot be replicated, researchers must be cautious about fishing through their findings. As Mills (1993, p. 1196) observed, "If you torture your data long enough, they will tell you whatever you want to hear."

13.9 BRACKETING ESTIMATES OF EFFECTS

No research design will be free from all threats to internal validity. Even when all such threats have been addressed, there will always be uncertainty about whether the threats have been addressed properly. There will be uncertainty, if for no other reason, than that statistical procedures rest on assumptions, and there will always be some doubt about whether the assumptions are correct. There will also be uncertainty about whether design features (such as methods of design elaboration) have perfectly removed the effects of threats to internal validity.

As a result, it is unreasonable to think that a point estimate (i.e., a single number) will be exactly equal to an effect size. The best a researcher can do in most cases will be to estimate an effect size within a range of values. Ideally, the range of scores brackets the size of the treatment effect, so one end of the bracket is likely smaller than the true effect size and the other end of the bracket is likely larger than the true effect size (Campbell, 1969a; Manski & Nagin, 1998; Reichardt, 2000; Reichardt & Gollob, 1986, 1987; Rosenbaum, 1987; Shadish, 2002; Shadish et al., 2002). In this way, a researcher can be confident that the effect size lies within the bracket.

A confidence interval is a special case of a bracket. It estimates a treatment effect within a range of scores as is required to bracket the size of a treatment effect. But a confidence interval takes account of uncertainty only due to chance differences, assuming the underlying statistical analysis is correct. Another way to say the same thing is that confidence intervals take account of uncertainty about an effect side due to chance only if the assumptions underlying the statistical procedures are perfectly correct. As just noted, however, there will usually (if not always) be uncertainty about assumptions. The solution is to use broader brackets—that is, broader lower and upper bounds on estimates of effects—than is provided by a confidence interval alone.

To take account of uncertainty about the assumptions that underlie statistical procedures and thereby better bracket the size of a treatment effect, researchers should

use a range of assumptions. In particular, researchers should try to impose statistical assumptions that would lead to an underestimate of the treatment effects and assumptions that would lead to an overestimate.

Researchers can also use design features (such as using more than one comparison group when implementing the methods of design elaboration) to create appropriate brackets of treatment effects. Studies of the Salk vaccine provide an example (Meier, 1972). The polio vaccine was investigated using both a large randomized experiment, and a nonequivalent group design. The nonequivalent group design used a single experimental group and two comparison groups. Specifically, the vaccine was given to second graders, and the results were compared to the results from both first and third graders who did not receive the vaccine. The design is diagrammed as:

$$\begin{array}{lll} \text{NR:} & & O_{\text{grade1}} \\ \text{NR:} & X & O_{\text{grade2}} \\ \text{NR:} & & O_{\text{grade3}} \end{array}$$

To the extent that age is related to the likelihood of contracting polio the effects of age were taken into account within a bracket. In particular, the comparison between the second graders and the first graders is biased by age in one direction, while the comparison between the second graders and third graders is biased by age in the opposite direction. That is, to the extent that age is a threat to internal validity, one comparison underestimates the effect of age and the other overestimates it. So the design takes account of the biasing effects of age by bracketing its effect between the two estimates.

Millsap, Goodson, Chase, and Gamse (1997; see Shadish et al., 2002) used bracketing to assess the effects of a school program on student achievement. Twelve schools that received the treatment were compared with 24 comparison schools. Each of the 12 experimental schools was matched with two comparison schools. One comparison school in each matched triplet had lower pretest achievement than its matched experimental school, and one comparison school had higher pretest achievement, thereby attempting to bracket the effects of selection differences.

Bitterman (1965) provided yet another example of using multiple comparison groups (see Campbell, 1969a; Reichardt, 2000; Rosenbaum, 1987, 2017) to bracket the effects of treatments (a method Bitterman called "control by systematic variation"). The theoretical question of concern was whether fish learn more slowly than rats. An alternative explanation for any differences in learning between fish and rats is differences in motivation due to differences in hunger. It was not known how to equate fish and rats in their hunger, but it was possible to vary the level of hunger so widely that it would bracket the possible range of hunger. The results were that fish showed no improvement in learning across the widest possible variation in hunger, so differences in hunger were ruled out as a plausible alternative explanation. Lee and Card (2008),

Manski and Nagin (1998), Shadish (2002), Reichardt and Gollob (1987), Rosenbaum (1987, 2017), Sagarin et al. (2014), and West and Sagarin (2000) provide further discussions and examples of bracketing.

Unfortunately, it will usually be difficult to be confident that a range of estimates properly brackets the size of a treatment effect. When not enough is known to feel confident about setting lower and upper bounds, researchers should still report the results of multiple analyses that rest on different sets of plausible assumptions. On the one hand, to the extent that the results agree, researchers can have increased confidence that the results do not rest tenuously on a single set of assumptions. On the other hand, if the results disagree, the researcher must be even more circumspect in reaching conclusions. In either case, the full set of results should be presented so that readers can judge the credibility of the range of treatment effect estimates for themselves. Researchers must be able to tolerate uncertainty and be honest about the degree of uncertainty that exists in a set of results. In addition, researchers should follow Cochran's (1965) advice and include a "Validity of the Results" section in their reports where the plausibility of assumptions and the likelihood that threats to internal validity remain are addressed honestly and openly.

13.10 CRITICAL MULTIPLISM

The principle of **critical multiplism** specifies that when in doubt about what course of action to take, multiple approaches should be used, all of which have different strengths and weaknesses (Cook, 1985; Patry, 2013; Shadish, 1993). I have already invoked the principle of critical multiplism under the guise of unfocused elaboration, where I have encouraged researchers to take account of threats to validity by using multiple estimates of effects that suffer from different threats to validity. I have also invoked the principle of critical multiplism under the guise of bracketing treatment effect estimates, where I have encouraged researchers to perform multiple statistical analyses based on different underlying assumptions. I emphasize critical multiplism again here because it is an overarching principle that applies even more broadly. When in doubt about the best way to operationalize a measure, the best approach is to use multiple operationalizations. When in doubt about which form of focused elaboration is best, the best approach is to use multiple forms of focused elaboration. When in doubt about the proper assumptions for statistical analyses, the best approach is to perform multiple analyses with a range of assumptions. When in doubt about the best methods for bracketing an effect size, the best approach is to use multiple bracketing methods. When in doubt about how to best interpret results, the best approach is to provide multiple interpretations. When in doubt about how to best customize a design for a given research setting, the best approach is to use multiple designs. Although randomized experiments are often superior to quasi-experiments, a randomized experiment combined with a quasi-experiment is often superior to a randomized experiment alone.

Of course, it is not possible to implement critical multiplism completely. Resources and time are limited. The point is simply that critical multiplism is the goal toward which to strive. It is unrealistic to believe that a treatment effect can be estimated perfectly. The proper perspective to take is that of using multiple approaches to triangulate on an effect size (Denzin, 1978). We validate results by their convergence. As already noted, to the extent that multiple methods and perspectives converge on the same results, we can have increased (but never complete) confidence in those results. To the extent that multiple methods and perspectives fail to converge on the same results, we must have decreased confidence in any one of the results.

Our methods are fallible. We cannot realistically hope to implement the perfect study. We must always be tentative in our conclusions because we always fall short of the ideal of critical multiplism and convergence of results.

13.11 MEDIATION

As has been noted throughout this volume, it can be a major accomplishment to obtain credible and precise estimates of treatment effects. Nonetheless researchers and other stakeholders would often like to know even more than estimates of treatment effects. They often want to know the processes or mechanisms by which treatment effects come about. Asking about the processes or mechanisms by which treatment effects come about is asking about the variables that mediate the effects of treatments on the outcome (Hayes, 2018; Judd, Yzerbyt, & Muller, 2014; MacKinnon, Cheong, & Pirlott, 2012; Mayer, Thoemmes, Rose, Steyer, & West, 2014; Preacher, 2015). In other words, a mediator (M) is a variable that causally comes in between the treatment (T) and the outcome (Y). More specifically, a mediator is influenced by the treatment, which in turn influences the outcome. This means the effect of the treatment on the outcome runs, at least partly, through the mediator. For example, the theory of cognitive dissonance specifies that, under certain conditions, performing an unpleasant task can cause a change in a person's attitudes about the task because of cognitive dissonance. That is, under the right circumstances, a treatment (T) that consists of performing an unpleasant task causes cognitive dissonance (M), which in turn causes a change in attitudes (Y). Alternative processes beside cognitive dissonance are also possible. For example, Bem (1972) proposed self-perception as an alternative (or additional) mediator by which performing an unpleasant task could produce a change in attitudes. Or to use the example of Mayer et al. (2014), an educational intervention (T) might influence academic achievement (Y) at least partly because it increases study time (M). Or a treatment (T) might reduce aggression (Y) by increasing social skills (M).

Understanding mediating processes can be important in both basic and applied research. In basic research, understanding **mediation** helps advance theories about the operation of treatment effects. In applied research, understanding mediation can help researchers improve the effectiveness and generalizability of ameliorative interventions.

Estimating treatment effects without assessing mediation is said to investigate the treatment as if it were a black box—with unknown contents. However, assessing mediation adds another level of complexity to assessing treatment effects. Although researchers may not be able to assess mediation with as much credibility as they can assess black-box treatment effects, they can try to look inside the black box when circumstances permit.

Mediation is not a topic usually addressed in expositions of quasi-experimentation, but, for an alternative perspective and approach, I include a few words about it here. I will introduce the classic way to assess mediation but see Imai, Keele, Tingley, and Yamamoto (2011).

The classic way to understand mediation is to use a path diagram such as that shown in Figure 13.2, which displays a very simple model of mediation. In practice, more complex models would be fit to the data where autoregressive effects of the variables are included (Cole & Maxwell, 2003; Maxwell & Cole, 2007; Maxwell, Cole, & Mitchell, 2011; Reichardt, 2011a). In Figure 13.2's diagram, arrows indicate causal effects. The arrows in the figure specify that the treatment (*T*) has an effect on both the mediator (*M*) and the outcome (*Y*). In addition, the mediator (*M*) is specified to have an effect on the outcome (*Y*). (The letters *a, b,* and *c* that are linked to the arrows reveal the size of these effects, and the letters *u* and *v* represent residuals). So the figure specifies that *T* has an effect on *Y* via two paths. The first path is from *T* to *Y* without going through *M,* which is called the direct effect of *T* on *Y*. This effect has size *c,* which means that a one-unit change in the treatment (which is the difference between the comparison and treatment conditions) has an effect on *Y* of size *c* units, while holding *M* constant. In the second path from *T* to *Y,* the treatment *T* has an effect on the outcome *Y* via the mediator *M*. This path is called the indirect effect of *T* on *Y*. This path has two parts. The part from *T* to *M* has size *a*. This means that the effect of the treatment (compared to the comparison condition) is to change *M* by the amount of *a* units. The part of the path from *M* to *Y* has size *b*. This means that a change in *M* of one unit causes a

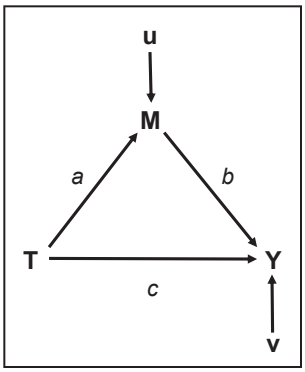


FIGURE 13.2. Path diagram for a simple model of mediation among the treatment (*T*), mediator (*M*), and outcome variables (*Y*).

change in *Y* of *b* units. So the size of the effect of *T* on *Y* via *M* has size *a* times *b* (which is specified as *ab*).

The direct and indirect effects taken together are called the total effect of *T* on Y. The total effect has size *c* + *ab*. The *c* portion is due to the direct effect of *T* on *Y,* and the *ab* portion is due to the indirect effect of *T* on *Y* via M. The paths *a, b,* or *c* can be any size. If, on one hand, either *a* or *b* is zero, there is no indirect effect of *T* on *Y* via *M*. In that case, the total effect of *T* on *Y* is due solely to the direct effect. On the other hand, if *c* is zero, there is no direct effect of *T* on *Y*. The total effect of *T* on *Y* is due solely to the indirect effect.

The path diagram in Figure 13.2 can be translated into two equations:

$$M_i = \alpha_M + (a\ T_i) + u_i \tag{13.1}$$

$$Y_i = \alpha_Y + (b\ M_i) + (c\ T_i) + v_i \tag{13.2}$$

In these equations, the α parameters are intercepts, which are usually of little interest. The other terms are as defined above. The values of *a, b,* and *c* can be estimated by fitting these equations using two separate regressions, or they can be estimated by fitting the two equations simultaneously, using structural equation modeling. Calculating standard errors for estimates of *a, b,* and *c* is straightforward (which allows the researcher to conduct statistical significance tests or create confidence intervals). Calculating a standard error for the product of *ab* (which is the indirect effect) is more complicated because, even if the estimates of *a* and *b* are each distributed normally, the estimate of the product of *a* times *b* will not be (MacKinnon, Cheong, & Pirlott, 2012).

The total effect of *T* on *Y* can be estimated by adding the estimates of *ab* to *c* from Equations 13.1 and 13.2. Alternatively, the total effect of *T* on *Y* can be estimated with the model:

$$Y_i = \alpha_T + (d\ T_i) + w_i \tag{13.3}$$

where *d* equals *ab* + *c* and where α_T and w_i are an intercept and error term, respectively.

The path diagram in Figure 13.2 includes a single mediating variable, *M*. It is possible to have more than one mediator and include them in the model as well. The inclusion of additional mediators follows the same logic. See Shadish et al. (2002, p. 186) for an example involving a series of potential mediators.

To estimate a mediation effect, the mediating variable *M* must obviously be measured and included in Equations 13.1 and 13.2. In addition, the mediator *M* must come temporally in between the treatment (*T*) and the outcome (*Y*). Measuring *M* at different time lags between *T* and *Y* (e.g., either closer in time to *T* or closer in time to *Y*) could produce different estimates of *a, b,* and *c* (Reichardt, 2011a).

Now consider how the values of the parameters in Equations 13.1 to 13.3 are to be estimated, which is the purpose of mediation analysis. First, consider the task of

estimating mediation effects in the context of a between-groups randomized experiment where the treatment conditions (represented by the *T* variable) are randomly assigned to participants. In such a randomized experiment, with perfect compliance and no missing data, Equations 13.1 and 13.3 could be fit to the data using ordinary least squares to produce unbiased estimates of *a* and *d*. The lack of bias is a result of the random assignment of treatment conditions. However, the estimates of *b* and *c* in Equation 13.2 are not guaranteed to be unbiased. That is, because *M* is not randomly assigned to participants, *M* might be related to the residual *v* in Equation 13.2, and that would lead to biases in the estimate of both *b* and *c*. For example, biases can arise if variables, not included in the equations, cause both *M* and *Y* (see Mayer et al., 2014). Biases can also arise if another mediator, not included in the model, intervened between *X* and *Y*. If that omitted mediator were correlated with the included mediator (*M*), the values of both *b* and *c* would be biased. For example, if the omitted mediator were positively correlated with the included mediator and had a positive effect on *Y,* the estimate of *b* could be positively biased, perhaps making the included mediator appear to have an effect when it did not. The point to be made is that the specification of Equation 13.2 might be incomplete in ways that could bias estimates of effects, even in a randomized experiment. To avoid such biases, the omitted variables would need to be included in the equation (Reichardt & Gollob, 1986). Because it is hard to know if all the relevant variables are included in the models, the results of any analyses to estimate the values of *b* and *c* must be interpreted with great caution. Indeed, estimating *b* and *c* must proceed with the same degree of caution as is necessary in estimating the treatment effect in the nonequivalent group design.

Second, consider the case of estimating mediation effects in the context of a nonequivalent group design. That is, assume the treatment (*T*) has not been assigned to participants at random. In that case, things are even worse than in the context of a randomized experiment. The estimates of none of the parameters in Equations 13.1–13.3 are guaranteed to be unbiased. Without random assignment of participants to treatments (*T*), estimating the effect of *T* on *M* (i.e., estimating the value of *a* in Equation 13.1) and estimating the total effect of *T* on *Y* (i.e., estimating the value of *d* in Equation 13.3) must be undertaken as if the design were a nonequivalent group design because that is what it is. The researcher must therefore use the methods of Chapter 7. The same difficulties and cautions hold for estimating the values of *b* and *c* in Equation 13.2, as in the case where the treatment (*T*) has been assigned to participants at random.

Another potential problem is measurement error in the variables. The treatment variable (*T*) is presumed to be measured without error, and any measurement error in *Y* introduces no bias (see Section 7.4.2). But measurement error in *M* or in other variables that would be added to the models as covariates does introduce bias. One way to cope with the effects of measurement error in mediators or other covariates is to use latent variable structural equation models as described in Section 7.4.2.

The conclusions to be drawn are the following. As noted in Chapter 4, random assignment to treatment conditions provides advantages in estimating the effects of

the treatment on outcomes. Even without either noncompliance or attrition, however, random assignment to treatment conditions is not sufficient to guarantee unbiased estimates of mediation effects. Estimating mediation effects, even in a randomized experiment, requires the same cautions as estimating treatment effects in the nonequivalent group quasi-experiment, which are described in Chapter 7. In addition, without random assignment, mediation effects must be estimated as if the design is a quasi-experiment—because that is what it is.

13.12 MODERATION

It is useful to know the average effect of a treatment for given participants, times, settings, and outcome measures. It is also useful to know how the treatment effect varies across participants, times, settings, and outcome measures (Brand & Thomas, 2013). For example, researchers can ask whether an antidepressant medication works as well for patients with mild depression as for patients with severe depression (Fournier et al., 2010). Does an early childhood intervention work as well for girls as boys (Anderson, 2008)? Does an intervention for drunken driving work the same during daytime as nighttime hours? Does an effect decay over time? A study of behavior modification might ask if the intervention is as successful at home as at school. How much does an educational innovation improve test performance in reading as compared to writing skills? Such questions ask about the moderation of the treatment effect—which is the same as asking about treatment effect interactions. In several chapters, I have shown how to assess treatment interactions using the statistical methods I have described.

Besides assessing moderation, including interactions in an analysis can also serve to increase the power and precision of statistical results. For example, compared to ignoring an interaction, fitting an ANCOVA model allowing an interaction between a covariate and the treatment can not only reveal how the effect of the treatment varies with the covariate but can also increase the precision with which the average treatment effect is estimated. The interactions of treatment effects across participants, times, settings, and outcome measures can also be assessed by comparing the results across multiple studies, such as by using research synthesis (e.g., meta-analysis).

It is important to be cautious in studying interactions because of the potential to fish for results. For example, many covariates will often be available to assess interactions with multiple participant characteristics, and it would be easy to find interactions that appear to be large just by chance alone. Researchers need to remember the chance that at least one Type I error increases with each statistical significance test that is performed, unless corrections with Bonferroni alpha-level adjustments or other methods are applied (Benjamini & Hochbert, 1995; Benjamini & Yekutieli, 2001). Conversely, estimates of interactions often have less statistical power and precision than estimates of average effects, so confidence intervals can be wide in small samples. Do not then just estimate average effects; also assess interactions to reveal how effects vary across

participants, times, settings, and outcome measures—but be cautious in interpreting the results whether positive or negative.

13.13 IMPLEMENTATION

Researchers must attend to myriad pragmatic details involved in implementing a research project, if treatment effects are to be estimated credibly, whether the study is a randomized experiment or a quasi-experiment (Boruch et al., 2009; Chambless & Hollon, 2012; U.S. Department of Education, 2017). Many issues relevant to implementation are addressed in the preceding chapters and in the preceding principles. A brief description of a few implementation issues is also given here.

13.13.1 Intervention

Researchers need to ensure that the treatment, and not just the research study, is implemented properly (Century, Rudnick, & Freeman, 2010; Coalition for Evidence-Based Policy, 2003; Sechrest et al., 1979). A theory of the treatment intervention should guide the design of the study, especially in applied field settings. A theory of the intervention is often specified using a logic model that details how a program is intended to operate and the mechanisms by which the program is expected to achieve its intended results (W. K. Kellogg Foundation, 2004). Adequate resources must be available to implement the treatment as intended. Staff need to be adequately trained and sufficiently competent to implement the intended treatment with fidelity, given the available resources. In addition, the nature of the treatment as implemented needs to be documented.

13.13.2 Participants

The researcher must gauge the number of participants needed for adequate power and precision in the estimation of treatment effects and ensure that an adequate supply of participants is available for the study. A Consolidated Statement of Reporting of Trials (CONSORT) flowchart can be used to assess the almost inevitable dwindling of participants at each stage of the research study from initial eligibility, to acceptance of the invitation to participate in the study, to completion of initial pretests, to compliance with treatment protocols, and to completion of posttest measures (Altman et al., 2001; Mohler, Schultz, & Altman, 2001). The researcher should ensure that participants meet all the necessary eligibility requirements. Participants must be tracked to ensure they complete the treatment and measurements. The researcher should document the characteristics of both the participants and the potential participants who drop out of the study. If participants are intended to be randomly assigned to treatment conditions, the researcher should use procedures that ensure that the randomization protocol is properly followed (Boruch & Wothke, 1985; Braucht & Reichardt, 1993).

13.13.3 Times and Settings

The times and settings of a study must be described accurately because they may be relevant to the success of the treatment. For example, a job training program may be more successful when implemented in times of economic expansion than in times of economic contraction. Stakeholders also need to know if the treatment was implemented to assess its effectiveness under typical conditions or if the treatment was implemented to assess its efficacy under ideal conditions.

13.13.4 Measurements and Statistical Analyses

Pretest and posttest measures need to be selected or crafted so that they are sufficiently sensitive to detect treatment effects. The reliability and validity of the measurement instruments must be assessed. Researchers need to consider collecting measures to assess undesirable effects, to detect both short- and long-term treatment effects, and to document costs. Researchers need to ensure that adequate measures are available to assess initial selection differences, treatment compliance, attrition, and statistical interactions. Researchers also need to collect whatever data might be available to test the assumptions that underlie statistical analyses. And they need to have the requisite expertise to design studies and implement statistical analyses.

13.14 QUALITATIVE RESEARCH METHODS

Qualitative research methods are research strategies that rely on rich qualitative accounts of participants' and others' perspectives, usually in the form of open-ended data collection procedures such as in-depth interviews (Glaser & Strauss, 1977; Reichardt & Rallis, 1994). In contrast to qualitative research methods, the design and analysis procedures described so far in this volume are considered quantitative research methods. Research that incorporates both quantitative and qualitative research methods is often called mixed-methods research (Creswell & Plano Clark, 2007). Following Harding and Seefeldt (2013), qualitative methods could be combined with the quantitative methods described in this volume to enhance quasi-experimentation in a variety of ways, including the following.

Qualitative methods (such as in-depth interviews) can be used to elucidate both the nature of the treatment conditions received by the participants and the ways the participants interpret and experience the treatment conditions. Through the use of qualitative methods, service providers and participants can provide insights into the way the treatment was delivered and the fidelity of its delivery. Qualitative methods can enlist service providers and participants in helping to discern which components of a multifaceted treatment are most responsible for its effects. Qualitative methods can also be used to identify potential mechanisms by which treatments have their effects.

Qualitative methods can be used to elucidate the processes by which participants are assigned or selected into treatment conditions. Qualitative interviews with whomever is responsible for assignment to treatment conditions (either participants or administrators) could illuminate the variables that need to be collected to adequately model selection, such as is required in propensity score analysis. For example, qualitative methods (such as in-depth interviews) could be used to discover the forces or decision-making processes that lead participants in nonequivalent group designs to select one treatment rather than another. Qualitative methods could be used to determine what makes some participants treatment compliers, while others are noncompliers. They could also be used to discover the ways in which scores on the quantitative assignment variable are being manipulated in regression discontinuity designs so as to understand (and perhaps eliminate) the biases otherwise produced. In addition, qualitative methods could be used to understand the reasons why some participants drop out of a study and thereby reduce such attrition.

Qualitative methods can be used to understand how quantitative pretest and posttest measures are interpreted by participants. Ambiguities in quantitative surveys and questionnaires can be uncovered using qualitative interviews. The degree to which participants take measurement seriously can be uncovered using qualitative methods. Qualitative methods can be used to discern the underlying constructs that are being assessed with quantitative measures and thereby help assess their validity and reliability.

Qualitative measures can be used to discover how and why treatment effects vary across different causes, participants, times, settings, and outcome measures. For example, a program to encourage employment by providing incentives and support such as child care, health insurance, and income subsidies had demonstrably different effects across participants, which were uncovered using qualitative methods (Harding & Seefeldt, 2013). The program had no effects on two distinguishable groups of participants, either because the barriers to employment those participants faced were greater than could be overcome by the program or because no substantial barriers were present for those participants. Only a middle group, which experienced barriers addressed by the program, showed improvements. The differences among these three groups were discovered only through in-depth longitudinal interviews with a random subsample of 43 families. Or consider a study of the effects of treatments that included after-school subsidies on behavior and academic achievement (Harding & Seefeldt, 2013). Treatment effects were found for boys but not for girls. Only qualitative interviews revealed the reason. Parents were devoting more program resources to their sons than to their daughters because the parents were more worried about the difficulties their sons experienced (due to gangs and violence), compared to those their daughters experienced, in their low-income environments.

In sum, qualitative methods can fill in many of the gaps in understanding that often arise in quantitative research. Together, qualitative and quantitative methods provide deeper insights than either can provide by themselves.

13.15 HONEST AND OPEN REPORTING OF RESULTS

Researchers should report the results of research honestly and openly. That means reporting the truth, the whole truth, and nothing but the truth. Like anyone, the researcher can be susceptible to omitting the whole truth. Researchers should not just report the results that support their hypotheses; they should also report results that fail to confirm, or even oppose, their hypotheses.

Researchers also need to be open to admitting the limitations of their studies. Some questions will always be left unanswered. Threats to validity will always remain. As already noted, Cochran (1965) recommends that all empirical research publications include a section entitled "Validity of the Results" in which researchers forthrightly acknowledge the uncertainties that remain in a research study. Researchers should make suggestions for how to conduct future research to address unanswered questions and remove ever present uncertainties (because, as noted, uncertainty is best reduced by accumulating the results of multiple studies). Editors of journals need to accept research for publication without penalty because researchers openly admit to the inevitable limitations and uncertainties that exist even in high-quality research. To avoid the file drawer problem, editors need to be willing to publish high-quality research reports, even if their estimates of treatment effects do not reach statistical significance. High-quality research that reports null results but has treatment effect estimates with narrow confidence intervals can deserve publication just as much as, if not more, than research with statistically significant results but with treatment effect estimates that have wide confidence intervals. Null results with narrow confidence intervals can alert stakeholders to treatment effects that are likely small, which can be just as important as alerting stakeholders to treatment effects that are likely large.

13.16 CONCLUSIONS

Estimating treatment effects credibly is not an easy task, especially in field settings. Much can go wrong. Nonetheless, the task of estimating effects, especially in field settings, is an important one that we should try to do as well as we can. If we are to test theories of behavior under realistic conditions, we need to estimate effects in field settings. If we are to ameliorate the numerous social problems that plague our society we must estimate treatment effects in field settings. If we are to estimate treatment effects credibly in field settings, we need to have the most powerful tools at our disposal. Those tools include quasi-experiments as well as randomized experiments. I have described the logic of both randomized and quasi-experiments. I have also presented general principles for implementing those designs and analyses to produce the most credible estimates of treatment effects possible.

13.17 SUGGESTED READING

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton, NJ: Princeton University Press.

—Provide an overview of the methods of data analysis from a range of quasi-experimental designs.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York: Cambridge University Press.

—Gives a more advanced treatment of statistical analysis for causal inference in both randomized and quasi-experiments.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford University Press.

—Provides a detailed accounting of the use of both randomized experiments and quasi-experiments in the context of educational research.

Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Cambridge, MA: Harvard University Press.

—Provides a very readable overview as well.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

—Required reading for further insights about quasi-experimentation.

Appendix

The Problems of Overdetermination and Preemption

A.1 THE PROBLEM OF OVERDETERMINATION

In agreement with some other philosophers, Scriven (2008, p. 16) has stated that a counterfactual approach to causality is "not the correct analysis of causation" because of the **problem of overdetermination** (also see Scriven, 2009; Cook et al., 2010). The problem of overdetermination arises when two causes simultaneously produce the same effect. For example, consider a criminal who is put to death by a firing squad of five shooters. If the shooters shoot simultaneously, there is no way to determine, according to the overdetermination criticism of the counterfactual definition, which of the five was responsible for the death. For example, it cannot be said that any one shooter was responsible for the death because if any one of them had not fired, the criminal would still have died. So even though the criminal was killed by the shooters, none of the shooters is responsible for the death, according to the criticism. Nor, so it is said, does the counterfactual definition permit you to say that all five of the shooters caused the death. That's because death would have resulted if only one of the shooters had fired rather than if all five had. The point is that the counterfactual definition, according to the overdetermination criticism, cannot decide who is responsible for the death of the criminal. In other words, in the face of overdetermination, the causal question cannot be answered unambiguously using a counterfactual definition.

This problem of overdetermination is a problem for the Cause Question rather than for the Effect Question (see Section 2.5). That is, the overdetermination dilemma arises for the question of who is responsible for the death: What is the cause of a given effect that is the Cause Question? When the Effect Question is asked instead of the Cause Question, no problem of overdetermination arises. In other words, no ambiguity arises when we ask what is the effect of the shooters—as long as both the treatment and comparison conditions are well specified, as I have argued should be done in all applications (see Section 2.1). For example, if the treatment and comparison conditions are five shooters versus no shooters, then the death is unambiguously an effect of that difference. Unambiguous answers arise

for any other clearly specified counterfactual treatment comparison. Try another example: if the treatment conditions were to be specified as one shooter versus no shooters, then the death would again be unambiguously the effect of that difference. Or try yet another scenario: if the treatment conditions consist of five shooters versus four shooters, then that difference is again unambiguous—the death is not the effect of that difference, for death would have occurred under either condition. Another way to conceptualize the last scenario is the following. The difference between four and five shooters firing is the difference between one shooter firing and that shooter not firing in the context of four other shooters firing. In that context, one shooter unambiguously makes no difference. No matter which comparison of treatment conditions is drawn, an unambiguous answer can be given about the effect (which is the Effect Question). Thus, overdetermination is not a problem when one is asking the Effect Question, which is the focus of experimentation. That is all that is required for us to proceed with our counterfactual definition and our discussion of experimentation.

So Scriven is wrong if his comment is taken to imply that overdetermination is a problem for the counterfactual definition given in Section 2.1. Rather, the problem is not overdetermination. The problem is that the Effect Question is sometimes confused with the Cause Question (see Section 2.5). The Cause Question requires a definition of a cause for a given effect. That definition has long been a source of contention for philosophers, among others, and is subject to the criticism of overdetermination. Scriven (1968, 1975) has made important contributions to the explication of the cause of a given effect. However, a completely adequate definition of a cause for a given effect (the Cause Question) has proven to be elusive (Brand, 1976, 1979). In contrast, the effect for a given cause (the Effect Question) has a simple counterfactual definition, as I have shown (see Sections 2.1 and 2.3).

The bottom line is the following. When philosophers take up the topic of causality, they typically address the Cause Question rather than the Effect Question. Philosophers have long known about the problem of overdetermination for definitions of causes (the Cause Question). Consequently, it might not be possible to provide a satisfactory counterfactual definition of a cause for a given effect (the Cause Question). Note, however, that the counterfactual definition of an effect for a given cause (the Effect Question) that I have given is untroubled by overdetermination.

A.2 THE PROBLEM OF PREEMPTION

Philosophers have also raised the related **problem of preemption** as a potential stumbling block for a counterfactual explication of causality. In the overdetermination example given earlier, multiple shooters all fire at the same time. In preemption, in contrast, the shooters are set to fire one after the other, so that one preempts the other. For example, suppose two sharpshooters (A and B) are prepared to assassinate a despotic dictator C. If A does not shoot before a certain time, B will shoot a moment after that time elapses. Further, suppose A shoots right as the time expires and that A's shot kills C. Under these circumstances, C would die whether or not A shoots, so, it might seem, the counterfactual definition does not lead to the conclusion that A killed C. That is, the counterfactual definition is said to lead to the conclusion that A did not kill C because both the treatment (A shooting C)

and the comparison condition (A not shooting C) lead to the same outcome (C's death). Therefore, the treatment has no effect, which seems odd because A obviously did kill C. But that is not a correct interpretation of the counterfactual definition. The proper interpretation agrees with the commonsense notion that A did indeed kill C. It is important to note, however, that A's killing of C merely shortened C's life by a brief moment in time. To clarify the apparent (but not real) dilemma, compare the preceding scenario to the following one. Suppose shooter B had been present only 20 years later. That is, suppose A shoots and kills C but that if A had not shot C, C would have been shot by B 20 years later. Under these circumstances, we do not hesitate to say (and the counterfactual definition correctly specifies) that A killed C and shortened his life by 20 years. The only difference between the two hypothetical scenarios is the difference between a brief moment and 20 years. In both cases, the counterfactual definition of an effect provides the same (and the correct) answer that the result of A's shooting was C's death (even if A's shot shortened C's life by only a brief moment).

To summarize, perhaps a counterfactual definition of the cause of a given effect (the Cause Question) is threatened by the problems of overdetermination and preemption. However, the counterfactual definition of the effect of a given cause (the Effect Question) is not threatened by the apparent dilemmas of either overdetermination or preemption, contrary to what is sometimes suggested.

Glossary

Always-takers: Units (e.g., participants) that receive the treatment no matter the treatment condition to which they were originally assigned.

Analysis of covariance (ANCOVA): Statistical procedure that is an instance of ordinary least squares regression. Used for assessing differences between treatment groups by regressing outcome scores onto a variable representing treatment assignment and covariates.

Analysis of variance (ANOVA): Statistical procedure that is an instance of ordinary least squares regression. Used for assessing differences between treatment groups by regressing outcome scores onto a variable representing treatment assignment.

Applied research: Research aimed primarily at solving problems and improving the human condition.

Assignment based on an explicit quantitative ordering: Assignment where units of the most prominent size-of-effect factor (e.g., participants) are ordered along a quantitative dimension and assigned to treatment conditions based on a cutoff score on that dimension.

Attrition: When units (e.g., participants) drop out of a study before their participation is complete (e.g., before all the outcome measures have been collected). Differential attrition arises when the quantity and/or nature of units that drop out of a study differs across treatment conditions.

Autocorrelation: Time-series data are autocorrelated when observations over time are correlated rather than independent.

Autocorrelation function (ACF): The pattern of correlations between lagged time-series observations.

Autoregressive (AR) model: A model of autocorrelation in time-series data where current residuals are caused by one or more prior residuals.

Autoregressive moving average (ARMA) model: A model of autocorrelation in time-series data that can contain autoregressive and/or moving average components.

Auxiliary variables: Variables added to an analysis to convert otherwise missing not at random (MNAR) data into missing at random (MAR) data and thereby reduce bias in estimating treatment effects.

Average treatment effect (ATE): The average treatment effect across all units (e.g., participants).

Average treatment effect on the treated (ATT): The average effect of the treatment on the units (e.g., participants) that received the treatment. Also called the treatment-on-the-treated (TOT) effect.

Average treatment effect on the untreated (ATU): The average effect of the treatment on the units (e.g., participants) that did not receive the treatment, if they had received it.

Balance: When pretreatment measures have the same distributions in treatment and comparison conditions.

Basic research: Research aimed primarily at testing theories and improving knowledge of natural phenomena.

Between-groups comparison: A comparison to estimate a treatment effect where a group of units (e.g., participants) receives the treatment condition, while another group of units (e.g., participants) receives the comparison condition.

Between-groups randomized experiment: A randomized experiment in which units (e.g., participants) are randomly assigned to treatment conditions.

Blocking: Analysis procedure wherein units (e.g., participants) are placed into categories (i.e., blocks) based on their scores on one or more pretreatment measures. Also called stratification or subclassification.

Broken randomized experiment: Randomized experiment where some units (e.g., participants) do not comply with treatment assignment and/or have missing data on outcome measures.

Caliper distance: Size of the difference between matched units (e.g., participants) on scores on a covariate.

Causal function: Notion that an effect size (ES) is a function of the Cause (C), Participant (P), Time (T), Setting (S), and Outcome Measure (O).

Cause Question: What is a cause of a given effect?

Centering: Centering a variable around a given value means subtracting that value from the variable.

Chance differences: A threat to internal validity due to a grab bag of essentially random influences that do not rise to the level of specific history events, maturational changes, testing effects, instrumentation changes, and so on.

Change-score analysis: Statistical procedure where the dependent variable is the difference between the posttest and pretest scores.

Cluster-randomized experiment: Design where aggregates of units (e.g., participants) rather than individual units are assigned to treatment conditions at random, with data being available from the individual units as well as from the aggregates of the units.

Cohorts: Groups of units (e.g., participants) that are similar on some characteristic, often the age of the participants.

Comparative interrupted time-series (CITS) design: An interrupted time-series (ITS) design with an added comparison time series of observations.

Comparison across outcome measures: A comparison in which outcome measures are the size-of-effect factor that varies most prominently with the treatment conditions.

Comparison across participants: A comparison in which participants are the size-of-effect factor that varies most prominently with the treatment conditions.

Comparison across settings: A comparison in which settings are the size-of-effect factor that varies most prominently with the treatment conditions.

Comparison across times: A comparison in which times are the size-of-effect factor that varies most prominently with the treatment conditions.

Compensatory equalization of treatments: When administrators, or others, provide extra resources to the units (e.g., participants) in the comparison condition to compensate for the advantages provided by the experimental treatment.

Compensatory rivalry: Arises when people who are assigned to the comparison condition perform better than they would have otherwise because they are aware that other people received a more desirable treatment.

Complier average causal effect (CACE): The average treatment effect among units (e.g., participants) that complied with the original treatment assignment. Also called the local average treatment effect (LATE).

Compliers: Units (e.g., participants) that accept the treatment to which they are assigned.

Confound: Anything else, besides the treatment, which varies across the treatment and comparison conditions.

Construct validity: Is concerned with correctly labeling the five size-of-effect factors in a causal relationship.

Construct validity of the cause: Is concerned with whether the treatment and comparison conditions are properly labeled in the conclusions reached about an effect size.

Construct validity of the outcome measure: Is concerned with whether the outcome measures are properly labeled in the conclusions reached about an effect size.

Construct validity of the participants: Is concerned with whether the participants are properly labeled in the conclusions reached about an effect size.

Construct validity of the setting: Is concerned with whether the settings are properly labeled in the conclusions reached about an effect size.

Construct validity of the time: Is concerned with whether the times and time lags are properly labeled in the conclusions reached about an effect size.

Continuity restriction: Notion that there are no discontinuities in level or slope in the regression surfaces in the regression discontinuity design, except because of a treatment effect.

Counterfactual definition of a treatment effect: A definition of a treatment effect where only one of the outcomes (either the treatment or the comparison outcome) can arise in practice. The other outcome cannot arise and so is contrary to fact.

Covariate: A pretest measure that varies with the outcome and is added to the statistical analysis for purposes such as to remove bias, increase power and precision, or assess interaction effects.

Critical multiplism: Using multiple methods with different strengths and weaknesses so that, to the extent the results from the multiple methods agree, the credibility of results is increased.

Crossovers: Units (e.g., participants) assigned to the comparison condition that manage to obtain the treatment that otherwise is reserved for those assigned to the treatment condition.

Cyclical autocorrelation: When observations at similar cyclical time points—say, 12 months apart—are correlated rather than independent.

Cyclical change: A threat to internal validity where a pretreatment observation is collected during one part of a cycle, while a posttreatment observation is collected during a different part.

Defiers: Units (e.g., participants) that refuse whichever treatment they are assigned. If assigned to the treatment, they are no-shows who participate in the comparison condition. If assigned to the comparison condition, they cross over to the treatment condition.

Difference-in-differences (DID) analysis: Analysis where the treatment effect is estimated as a difference between treatment conditions in their differences between pretreatment to posttreatment observations.

Diffusion or imitation of treatments: When the treatment given to units (e.g., participants) in the treatment condition becomes available to units in the comparison condition.

Doubly robust method: An analysis in the nonequivalent group design that models both outcome scores and selection differences. If either model is correct, the estimates of treatment effects will be unbiased.

Dry-run analysis: Analysis to estimate a treatment effect using double pretests in a nonequivalent group design where the null hypothesis is known to be true. Statistically significant results lead the researcher to suspect hidden bias and vice versa. An instance of a falsification test.

Effect Question: What is an effect of a given cause?

Enroll-only-if-encouraged participants: Units (e.g., participants) who are the compliers with the treatment assignment in a randomized encouragement design. These units are nudged into the treatment by the encouragement but otherwise would not participate in the treatment.

Estimate-and-subtract method of design elaboration: Where the second estimate in focused design elaboration is an estimate of the effect of a threat to validity.

Estimation-maximization (EM) algorithm: A method for coping with missing data where missing data are imputed and a covariance matrix and means are recomputed iteratively to produce maximum likelihood estimates.

Exclusion restriction: With noncompliance, no-shows and crossovers would have the same outcome whether assigned to the treatment or comparison condition. With instrumental variables, the instrument is excluded from the model of the outcome. The instrumental variable has no direct effect on the outcome variable.

Experiment: Any empirical comparison used to estimate the effects of treatment; either a randomized experiment or a quasi-experiment.

External validity: Concerns the generalizability of the results of a study. Asks whether the researcher studied the five size-of-effect factors of interest and, if not, whether the results that were obtained generalize to the size-of-effect factors that are of interest.

Externalities: Where the effect of a treatment has unintended consequences outside the bounds of the intervention.

Factorial design: A design used to assess the way the effects of two or more treatments might interact to produce different effects when implemented together than when implemented separately.

Falsification test: A means to test the assumptions of an estimation strategy that is often a test to see if pseudo treatment effects are present.

Focal local comparison groups: Comparison groups that are like experimental groups in both location and pretreatment characteristics.

Focused design elaboration: Separate estimates are used to disentangle the treatment effect from the effect of a shared threat to internal validity.

Full information maximum likelihood (FIML) analysis: A method for coping with missing data where a likelihood function for the given data is specified and a treatment effect estimate is derived by choosing values for unknown parameters that maximize the likelihood function for the given data.

Fuzzy regression discontinuity design: The probability of participation in the treatment conditions in a regression discontinuity design is not uniformly 1 on the treatment condition side of the cutoff score and/or not uniformly 0 on the other side. There are either no-shows, crossovers, or both.

Hidden bias: Bias due to selection differences on unobserved variables that remains after bias due to selection differences on observed covariates has been taken into account in the statistical analysis.

Hierarchical linear models (HLM): Statistical models used to analyze data from multiple (hierarchical) levels of units simultaneously. A lower level of units (e.g., students) are nested in a higher level of units (e.g., classrooms).

History effects: External events that introduce threats to internal validity. Differential history effects arise when history effects differ across treatment conditions. Such differential events are sometimes called hidden treatments (Rubin, 2005) or a threat to internal validity due to selection by history.

Hot deck method: A single-imputation method of coping with missing data where missing values are replaced by other values, from the same variable, that arise in the dataset.

Ideal comparison: The comparison that defines a treatment effect but is impossible to obtain in practice.

Ignorability: When equating the treatment groups on observed covariates is sufficient to remove bias due to all selection differences. The same as unconfoundedness.

Independence assumption: Assumption that an instrumental variable must not be correlated with omitted variables or the error term in the model for the outcome variable.

Instrumental variable: A replacement for a variable that would lead to a biased estimate if used in the data analysis. The instrumental variable (or instrument) must be correlated with the variable for which it is an instrument and affect the outcome only through its influence on the variable for which it is an instrument.

Instrumental variables (IV) analysis: Usually implemented with a two-stage least squares procedure. In the first stage, the variable that is being instrumented for (e.g., treatment status with noncompliance data in a randomized experiment, treatment assignment in the nonequivalent group design, or treatment status in a

fuzzy regression discontinuity design) is regressed onto the instrumental variable to produce a predicted instrumented variable. In the second stage, the posttest scores are regressed onto the predicted instrumented variable to produce an estimate of the treatment effect.

Instrumentation effects: Changes in measurement instruments that raise threats to internal validity. Differential instrumentation (sometimes called a threat to internal validity due to selection by instrumentation) arises when the measurement instruments differ across treatment conditions.

Intention-to-treat (ITT) analysis: Analysis that compares those assigned to the treatment condition to those assigned to the comparison condition, regardless of treatment actually received. The same as the treatment-as-assigned analysis.

Intention-to-treat (ITT) estimate. The mean difference between the posttest scores in the treatment and comparison conditions, as participants were assigned to those conditions.

Interaction effects: Differential effects of a treatment across different participants, times, settings, or outcome measures. Also called moderator effects. Interaction effects also refer to how two treatments might have different effects when combined than when implemented separately.

Internal validity: A special case of construct validity that concerns the labeling (or mislabeling) of the size-of-effect factor of the cause due to confounds that would have been present in a comparison even if the treatment had not been implemented.

Interrupted time-series (ITS) design: A quasi-experimental design to estimate a treatment effect where a series of observations is collected before a treatment is introduced and is compared to a series of observations collected after the treatment is introduced.

Intraclass correlation (ICC): The proportion of total variance in outcome scores that is due to between-cluster variability in a hierarchical-linear-model design.

Kernel regression: Regression analysis where the regression surface is smoothed by weighting scores nearby (more than scores far away from) the point at which the regression line is being estimated.

Latent variable: Unmeasured construct that is the underlying cause of observed measurements.

Listwise deletion: A method for coping with missing data where a unit's (e.g., participant's) data are included in the statistical analysis only if the unit has complete data on all variables in the statistical model.

Local average treatment effect (LATE): The same as the complier average causal effect (CACE).

Logistic regression: Regression analysis when the outcome variable is categorical (e.g., dichotomous) rather than continuous.

Matched-pairs *t*-test: A test of statistical significance where units (e.g., participants) are matched before being compared on an outcome measure.

Matching: Analysis procedure wherein units (e.g., participants) from one treatment condition are matched with units from the other treatment condition based on their scores on their pretreatment measures.

Maturation: A threat to internal validity due to the natural process that occurs within the units (e.g., participants) because of the passage of time.

Mean substitution: A single-imputation method for coping with missing data where the missing values of a variable are replaced by the mean of the variable calculated with the available data on that variable.

Mediation: The processes or mechanisms by which treatment effects come about.

Meta-analysis: A statistical method for combining results from different studies to derive conclusions about both average effects across studies and how effects differ across studies.

Methods of design elaboration: An estimate of a treatment effect is combined with a second estimate to disentangle the effects of a threat to internal validity from the effect of the treatment.

Missing at random (MAR). Missing data where the missing scores depend on observed variables but not unobserved variables—so that missingness is unrelated to the values of the missing scores once observed variables are included appropriately in the statistical model.

Missing completely at random (MCAR). Missing data where missingness on a variable is not related either to the missing values of that variable or to any other variables in the analysis.

Missing data: Data that were intended to be collected but are not available.

Missing not at random (MNAR): Missing data where the probability that scores are missing is related to the missing values themselves (were they available), even after including observed variables in the statistical model.

Moderator effects: Occur when the effect of a treatment varies across values of another variable. The same as a treatment effect interaction.

Monotonicity assumption: With noncompliance, there are no participants who would always refuse the treatment to which they are assigned (i.e., defiers). With instrumental variables, the instrument cannot push some participants into treatment while it pushes others out.

Moving average (MA) model: A model of autocorrelation in time-series data where current residuals share an error term with one or more prior residuals.

Multiple-baseline (MBL) design: A comparative interrupted time-series (CITS) design with switching replications, especially with short time series used in the field of applied behavior analysis.

Multiple imputation (MI): A method for coping with missing data where (1) multiple copies of the data are generated with imputed data for the missing values, (2) treatment effect estimates are derived from each of the multiple copies, and (3) the multiple estimates are combined to create a single-treatment effect estimate and its estimated standard error.

Never-takers: Units (e.g., participants) that do not receive the treatment even when assigned to the treatment condition. They would be no-shows if assigned to the treatment condition.

No-shows: Units (e.g., participants) assigned to the treatment condition that do not, for one reason or another, receive the treatment.

Noncompliance: When units (e.g., participants) fail to receive the treatment to which they were assigned. Also called nonadherence to treatment conditions.

Nonequivalent assignment to treatment conditions: Assignment to treatment conditions that is neither random nor based on an explicit quantitative ordering of the units of the most prominent size-of-effect factor.

Nonequivalent dependent variable: Outcome variable that shares the effects of threats to validity but is free of the effects of the treatment.

Nonequivalent group design: A quasi-experimental design where treatment effects are assessed by comparing groups of participants who were not assigned to treatment conditions at random (nor according to an ordering of the participants on an explicit quantitative variable).

One-group posttest-only design: A design where a single group is exposed to a treatment and then assessed on an outcome measure. There is no pretest measure or comparison group.

One-to-many matching: Where each unit (e.g., participant) assigned to either the treatment or comparison condition is matched to multiple units that are assigned to the other condition.

One-to-one matching: Where each unit (e.g., participant) from the treatment condition is matched with one unit from the comparison condition.

Ordinary least squares (OLS) regression: The classic regression procedure where estimates of parameters are chosen by minimizing the sum of the squared estimated residuals.

Pairwise deletion: A method for coping with missing data where means and covariance matrices are calculated using all the nonmissing data that are available.

Panel data: Time-series data collected on multiple units (e.g., participants) where the same units are observed at each time period.

Parallel trends assumption: Assumption that the average change from pretest to posttest would be the same in the treatment and comparison conditions in the absence of a treatment effect.

Partial autocorrelation function (PACF): The pattern of partial correlations between lagged time-series observations holding intermediate observations constant.

Pattern matching: The process by which a treatment effect is estimated wherein the predicted patterns of results due to a treatment are compared to the predicted patterns of results due to threats to validity. To the extent that the pattern predicted by the treatment fits the data better than do the patterns predicted by threats to internal validity, a treatment effect is plausible.

Per-protocol analysis: Analysis that compares those who completed the treatment protocols as they were originally assigned and discards those who did not complete the treatment protocols as originally assigned.

Phase: A period of either treatment or no treatment in a single-case design.

Polynomial Terms: The quadratic polynomial term of X is the square of X, the cubic polynomial term of X is the cube of X, and so on.

Posttest-only between-groups randomized experiment: A between-groups randomized experiment in which units (e.g., participants) are assessed after the treatment is implemented but not before.

Pre-experimental design: A design for estimating an effect based on a nonempirical comparison.

Pretest–posttest between-groups randomized experiment: A between-groups randomized experiment in which units (e.g., participants) are assessed both before and after the treatment is implemented.

Pretest–posttest design: A quasi-experimental design where a treatment effect is estimated by comparing an outcome before a treatment is introduced to an outcome after the treatment is introduced.

Principle of parallelism: The principle that any design option that exists for any one of the four size-of-effect factors of participants, times, settings, or outcome measures exists for the other three as well. For example, if a researcher could draw a comparison across participants, a parallel comparison could be drawn (at least in theory) across times, settings, and outcome measures.

Problem of overdetermination: A problem for the definition of a cause where two potential causes simultaneously produce the same effect.

Problem of preemption: A problem for the definition of a cause where two potential causes are present and one preempts the other.

Prominent size-of-effect factor: The one size-of-effect factor of participants, times, settings, or outcome measures that varies most prominently with the treatment conditions in a practical comparison used to estimate a treatment effect.

Propensity score: The probability that a unit (e.g., participant) is assigned to the treatment condition (rather than the comparison condition) based on the unit's scores on a given set of covariates.

Pseudo treatment effect: An apparent (but spurious) effect that appears to be a treatment effect that arises in data in which no treatment effect is present. Also called a placebo outcome.

Quantitative assignment variable (QAV): The variable upon which participants are assigned to treatment conditions in a regression discontinuity design.

Quasi-experiment: An empirical comparison used to estimate the effects of treatments where units (e.g., participants) are not assigned to treatment conditions at random.

***R* squared:** The proportion of the variance of the dependent variable explained by the independent variables in a regression model.

Random assignment: Where units (e.g., participants) are assigned to different treatment conditions at random.

Random sampling: Where units (e.g., participants) are a sample drawn at random from a larger population.

Randomized clinical (or controlled) trial (RCT): A randomized experiment.

Randomized encouragement design: Units (e.g., participants) are encouraged (or prompted) to partake of the treatment, or not to partake of the treatment, at random.

Randomized experiment: An empirical comparison used to estimate the effects of treatments where treatment conditions are assigned to units (e.g., participants) at random.

Region of common support: The range of propensity scores over which there is overlap between the treatment groups.

Regression discontinuity design: A quasi-experimental comparison to estimate a treatment effect where participants are assigned to treatment conditions based on a cutoff score on a quantitative assignment variable.

Regression model: Statistical procedure used to estimate treatment effects in which outcome variables are regressed upon (i.e., predicted using) other variables, which can include pretreatment variables and variables indicating treatment assignment.

Regression toward the mean: When scores on a pretest are selected because they are particularly high or low, scores on a posttest will tend to return to a more typical or average level.

Relevance assumption: Assumption that an instrumental variable is sufficiently correlated with the variable for which it is serving as an instrument.

Resentful demoralization: When people who are assigned to the comparison condition perform worse because they perceive that others are receiving a more desirable treatment and so become demoralized over their relative disadvantage.

Rubin causal model: Model that defines a treatment effect as a counterfactual difference between potential outcomes and explicates the nature and role of mechanisms for selection into treatment conditions.

Seasonality: A threat to internal validity where a pretreatment observation is collected during one part of a cycle while a posttreatment observation is collected during a different part, especially when the cycle occurs over the course of a year. The same as cyclical changes.

Selection bias: Bias due to selection differences.

Selection differences: Differences between the units (e.g., participants) in different treatment conditions.

Sensitivity analysis: Analysis that assesses the degree to which estimates of a treatment effect are sensitive to bias.

Sharp regression discontinuity design: When the probability of participation in the treatment conditions in a regression discontinuity design is uniformly 1 on the treatment condition side of the cutoff score and 0 on the other side.

Single-case design (SCD): An interrupted time-series (ITS) design, especially with short time series used in the field of applied behavior analysis.

Single imputation: A method for coping with missing data where missing values are replaced with a single value to create a single complete dataset—as opposed to multiple imputation where multiple replacements are made to create multiple complete datasets.

Size-of-effect factors: The five factors that determine the size of a treatment effect: the treatment or cause (C), the participants (P) in the study, the times (T) at which the treatments are implemented and effects are assessed, the settings (S) in which the treatments are implemented and effects are assessed, and the outcome measures (O) upon which the effects of the treatment are estimated.

Stable-unit-treatment-value assumption (SUTVA): The assumption that when exposed to a specified treatment, the outcome produced by a participant in a study will be the same (1) regardless of how the participants in the study were assigned to treatment conditions and (2) regardless of the treatments received by the other participants in the study.

Statistical conclusion validity: Addresses two questions: (1) is the degree of uncertainty that exists in the estimate of a treatment effect correctly represented and (2) is that degree of uncertainty sufficiently small (e.g., is the estimate of the treatment effect sufficiently precise and is a test of statistical significance sufficiently powerful)?

Strong ignorability: When there is ignorability and each unit (e.g., participant) has a nonzero probability of being in either the treatment or comparison condition.

Structural equation models (SEM): Statistical models for estimating effects especially when there are multiple equations with multiple dependent variables. Useful for modeling latent variables and mediation effects.

Switching replications: Design feature where two treatment groups receive the same treatment but at different times. The group that does not receive the treatment the first time the treatment is introduced provides the counterfactual comparison for the group that receives the treatment the first time. The roles of the groups are switched by the time of the second introduction of the treatment.

Testing effects: A threat to internal validity due to having collected a pretreatment observation.

Threat to validity: An alternative explanation for a putative treatment effect.

Tie-breaking randomized experiment: A randomized experiment added to a regression discontinuity design where those with QAV scores in the middle of the distribution are assigned to treatment conditions at random.

Transition: A change from one phase to another in a single-case design.

Treatment-as-assigned analysis: The same as intention-to-treat analysis.

Treatment-as-received analysis: Analysis that compares those who received the treatment condition (even if they were not assigned to the treatment condition) to those who received the comparison condition (even if they were not assigned to the comparison condition).

Treatment effect: The difference in outcomes between what happened after a treatment is implemented and what would have happened had a comparison condition been implemented instead, assuming everything else had been the same.

Treatment effect interaction: When the effect of a treatment varies with the values of another variable (such as a covariate). Includes when two treatments are present and the effect of one treatment is either enhanced or diminished by the presence of the other treatment.

Treatment-on-the-treated (TOT) effect: The average effect of the treatment on the units (e.g., participants) that received it. Also called the average treatment effect on the treated (ATT).

Two-stage least squares (2SLS) regression: Where two stages of regression models (such as for instrumental variable analysis) are fit to the data together.

Unconfoundedness: When equating the treatment groups on observed covariates is sufficient to remove bias due to all selection differences. The same as ignorability.

Unfocused design elaboration: Separate estimates of the treatment effect that are subject to different threats to validity are used to enhance the credibility of the results.

Units of assignment to treatment conditions: The size-of-effect-factor that is most prominently assigned to treatment conditions. The units could be participants, times, settings, or outcome measures (see Chapter 10).

Variance inflation factor (VIF): How much the variance of an estimate of a model parameter is increased because of the presence of multicollinearity.

Vary-the-size-of-the-bias method of design elaboration: Where the second estimate in focused design elaboration has the same amount of treatment effect as the first estimate but a different amount of bias.

Vary-the-size-of-the-treatment-effect method of design elaboration: Where the second estimate in focused design elaboration equals a different amount of the treatment effect as the first estimate but the same amount of bias due to a specified threat to validity.

White noise: When time-series data are not autocorrelated.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science, 59,* 495–510.

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review, 91,* 1369–1401.

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analysis of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology, 62,* 488–499.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: SAGE.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22,* 207–244.

Aiken, L. S., West, S. G., Woodward, C. K., & Reno, R. R. (1994). Health beliefs and compliance with mammography screening recommendations in asymptomatic women. *Health Psychology, 13,* 122–129.

Algina, J., & Olejnik, S. F. (1982). Multiple group time-series design: An analysis of data. *Evaluation Review, 6,* 203–232.

Algina, J., & Swaminathan, H. (1979). Alternatives to Simonton's analyses of the interrupted and multiple-group time-series designs. *Psychological Bulletin, 86,* 919–926.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20,* 93–114.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: SAGE.

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 72–89). Thousand Oaks, CA: SAGE.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., . . . Lang, T. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134,* 663–694.

Anderson, A. J. B. (1989). *Interpreting data: A first course in statistics*. London: Chapman & Hall.

Anderson, M. (2008). Multiple inference and gender differences in the effect of early intervention:

A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association, 103,* 1481–1495.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association, 91,* 444–455.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economics Perspectives, 15,* 69–85.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away?: Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association, 110,* 1331–1344.

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology, 59,* 177–181.

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics, 67,* 648–660.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46,* 399–424.

Azar, B. (1994, November). Eating fat: Why does the brain say, "Ahhh"? *APA Monitor*, p. 20.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20,* 40–49.

Ball, S., & Bogatz, G. A. (1970). *The first year of Sesame Street: An Evaluation*. Princeton, NJ: Educational Testing Service.

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Barnow, B. S., Cain, G. C., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In E. W. Stromsdorfer & G. Farkas (Eds.), *Evaluation studies review annual* (Vol. 5, pp. 43–59). Newbury Park, CA: SAGE.

Barth, J. (1967). *The floating opera.* New York: Anchor Books.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York: Academic Press.

Bem, D. J., Wallach, M. A., & Kogan, N. (1965). Group decision making under risk of aversive consequences. *Journal of Personality and Social Psychology, 1,* 453–460.

Benjamini, Y., & Hochbert, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,* Series B (Methodological), *57,* 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annuals of Statistics, 29,* 1165–1188.

Bentler, P. M., & Woodward, J. A. (1978). A Head Start reevaluation: Positive effects are not yet demonstrable. *Evaluation Quarterly, 2,* 493–510.

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology, 6,* 191–208.

Berk, R. A., & Rauma, D. (1983). Capitalizing on nonrandom assignment to treatment: A

regression discontinuity evaluation of a crime control program. *Journal of the American Statistical Association, 78,* 21–27.

Bertrand, M., Duflo, E., Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics, 119,* 249–275.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal?: A field experiment on labor market discrimination. *American Economic Review, 94,* 991–1013.

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science, 1,* 31–49.

Biglan, A., Hood, D., Borzovsky, P., Ochs, L., Ary, D., & Black, C. (1991). Subject attrition in prevention research. In C. G. Luekefeld & W. Bukoski (Eds.), *Drug abuse prevention intervention research: Methodological issues* (NIDA Research Monograph No. 107, pp. 213–234). Washington, DC: U.S. Goverment Printing Office.

Bitterman, M. E. (1965). Phyletic differences in learning. *American Psychologist, 20,* 396–410.

Black, D. A., Galdo, J., & Smith, J. A. (2007). Evaluating the bias of the regression discontinuity design using experimental data. Retrieved from *www.researchgate.net/publication/228646499.*

Black, S. E. (1999). Do better schools matter?: Parental valuation of elementary education. *Quarterly Journal of Economics, 114,* 577–599.

Blalock, H. M., Jr. (1964). *Causal inference in nonexperimental research*. Durham: University of North Carolina Press.

Blitstein, J. L., Murray, D. M., Hannan, P. J., & Shadish, W. R. (2005). Increasing the degrees of freedom in existing group randomized trials: The df* approach. *Evaluation Review, 29,* 241–267.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8,* 225–246.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With application to a study of accelerated schools. *Evaluation Review, 27,* 3–49.

Bloom, H. S. (Ed.). (2005a). *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage Foundation.

Bloom, H. S. (2005b). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York: Russell Sage Foundation.

Bloom, H. S. (2008). The core analytics of randomized experiments for social research. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 115–133). Thousand Oaks, CA: SAGE.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*(1), 43–82.

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 173–235). NY: Russell Sage Foundation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29,* 30–59.

Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology, 38,* 37–72.

Boruch, R. F. (1975). Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods and Research, 4,* 31–53.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: SAGE.

Boruch, R. F. (Ed.). (2005). *Place randomized trials: Experimental tests of public policy* [Special issue]. *Annals of the American Academy of Political and Social Science, 599.*

Boruch, R. F., & Foley, E. (2000). The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In L. Bickman (Ed.), *Validity and experimentation: Donald Campbell's legacy* (pp. 193–238). Thousand Oaks, CA: SAGE.

Boruch, R. F., & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluation. *Professional Psychology, 8,* 411–434.

Boruch, R. F., McSweeny, A. J., & Soderstrom, E. J. (1978). Randomized field experiments for program planning, development, and evaluation. *Evaluation Quarterly, 2,* 655–695.

Boruch, R. F., Weisburd, D., Turner, H. M., III, Karpyn, A., & Littell, J. (2009). Randomization controlled trials for evaluation and planning. In L. Bickman & D. J. Rog (Eds.), *The SAGE handbook of applied social research methods* (2nd ed., pp. 147–181). Thousand Oaks, CA: SAGE.

Boruch, R. F., & Wothke, W. (1985). Seven kinds of randomization plans for designing field experiments. In R. F. Boruch & W. Wothke (Eds.), *Randomization and field experimentation* (New Directions for Program Evaluation No. 28, pp. 95–113). San Francisco: Jossey-Bass.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Brand, J. E., & Thomas, J. S. (2013). Causal effect heterogeneity. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 189–213). Dordrecht, the Netherlands: Springer.

Brand, M. (Ed.). (1976). *The nature of causation*. Urbana: University of Illinois Press.

Brand, M. (1979). Causality. In P. D. Asquith & H. E. Kyburg, Jr. (Eds.), *Current research in philosophy of science*. East Lansing, MI: Philosophy of Science Association.

Braucht, G. N., & Reichardt, C. S. (1993). A computerized approach to trickle-process, random assignment. *Evaluation Review, 17,* 79–90.

Braucht, G. N., Reichardt, C. S., Geissler, L. J., Bormann, C. A., Kwiatkowski, C. F., & Kirby, M. W., Jr. (1995). Effective services for homeless substance abusers. *Journal of Addictive Diseases, 14,* 87–109.

Brewer, M. B., & Crano, W. D. (2014). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 11–26). New York: Cambridge University Press.

Briggs, D. C. (2004). Causal inference and the Heckman model. *Journal of Educational and Behavioral Statistics, 29,* 397–420.

Bryk, A. S., & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics, 1,* 127–155.

Buddelmeyer, H., & Skoufias, E. (2004). *An evaluation of the performance of regression discontinuity design on PROGRESA* (World Bank Policy Research Working Paper No. 3386; IZA Discussion Paper No. 827). Washington, DC: World Bank.

Cahan, S., & Davis, D. (1987). A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal, 24,* 1–12.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22,* 31–72.

Campbell, D. T. (1957). Factors relevant to validity of experiments in field settings. *Psychological Bulletin, 54,* 297–312.

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 81–106). New York: Holt, Rinehart & Winston.

Campbell, D. T. (1969a). Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 351–382). New York: Academic Press.

Campbell, D. T. (1969b). Reforms as experiments. *American Psychologist, 24,* 409–429.

Campbell, D. T. (1975). Degrees of freedom and the case study. *Comparative Political Studies, 8,* 178–193.

Campbell, D. T. (1978). Qualitative knowing in action research. In M. Brenner, P. Marsh, & M. Brenner (Eds.), *The social contexts of method* (pp. 184–209). London: Croom Helm.

Campbell, D. T. (1984). Forward. In W. M. K. Trochim, *Research design for program evaluation: The regression-discontinuity approach* (pp. 15–43). Beverly Hills, CA: SAGE.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis*. (New Directions for Program Evaluation, No. 31, 67–77.) San Francisco: Jossey-Bass.

Campbell, D. T. (1988). Can we be scientific in applied social science? In E. S. Overman (Ed.), *Methodology and epistemology for social science: Selected papers* (pp. 315–333). Chicago: University of Chicago Press.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate: Vol. 3. Disadvantaged child* (pp. 185–210). New York: Brunner/Mazel.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Campbell, D. T., & Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review, 3,* 33–53.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Cappelleri, J. C. (1991). *Cutoff-based designs in comparison and combination with randomized clinical trials*. Doctoral dissertation, Cornell University, Ithaca, NY.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review, 18,* 141–152.

Cappelleri, J. C., & Trochim, W. M. K. (2015). Regression discontinuity design. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed., Vol. 20, pp. 152–159). Amsterdam, the Netherlands: Elsevier.

Cappelleri, J. C., Trochim, W. M. K., Stanley, T. D., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review, 15*(4), 395–419.

Card, C., & Krueger, A. (1994). Minimum wage and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review, 84,* 772–784.

Carr-Hill, R. (1968). The concept of cause in criminology: A comment. *Issues in Criminology, 3,* 167–171.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31,* 199–218.

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods, 21,* 427–445.

Chambless, D. L., & Hollon, S. D. (2012). Treatment validity for intervention studies. In H. Cooper (Ed.), *The APA handbook of research methods in psychology* (pp. 529–552). Washington, DC: American Psychological Association.

Christianson, L. B. (1988). *Experimental methodology* (4th ed.). Boston: Allyn and Bacon.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct—recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58,* 1015–1026.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Coalition for Evidence-Based Policy. (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. Retrieved from *www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf*.

Coalition for Evidence-Based Policy. (2014). Which comparison-group ("quasi-experimental") study designs are most likely to produce valid estimates of a program's impact?: A brief overview and sample review form. Retrieved from *http://coalition4evidence.org/wp-content/uploads/2014/01/Validity-of-comparison-group-designs-updated-January-2014.pdf*.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13,* 261–280.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society,* Series A, *128,* 234–266.

Cochran, W. G. (1968a). Errors of measurement in statistics. *Technometrics, 10,* 637–666.

Cochran, W. G. (1968b). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24,* 295–313.

Cochran, W. G. (1969). The use of covariance in observational studies. *Journal of the Royal Statistical Society,* Series C, *18,* 270–275.

Cochran, W. G. (1972/2015). Observational studies. In T. A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor.* Ames: Iowa State University Press. (Reprinted in *Observational Studies, 1,* 126–136.)

Cochran, W. G., & Cox, G. (1957). *Experimental designs*. New York: Wiley.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: Indian Journal of Statistics, Series A, 35,* 417–446.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1985). *Applied multiple regression and correlation analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Routledge.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112,* 558–577.

Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive missing-data strategies in modern missing data procedures. *Psychological Methods, 6,* 330–351.

Conner, R. F. (1977). Selecting a control group: An analysis of the randomization process in twelve social reform programs. *Evaluation Quarterly, 1,* 195–244.

Cook, T. D. (1985). Post-positivist critical multiplism. In R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21–62). Beverly Hills, CA: SAGE.

Cook, T. D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (DHHS Publication No. PHS 90-3454, pp. 9–31). Rockville MD: U.S. Department of Health and Human Services.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24,* 175–199.

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *Annals of the American Academy of Political and Social Science, 589,* 114–149.

Cook, T. D. (2005). Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. *Annals of the American Academy of Political and Social Science, 599,* 176–198.

Cook, T. D. (2008a). Randomized experiments in education: Assessing the objections to doing them. *Economics of Innovation and New Technology, 16,* 31–49.

Cook, T. D. (2008b). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 142,* 636–654.

Cook, T. D. (2014). Generalizing causal knowledge in the policy sciences: External validity as a task of both multiattribute representation and multiattribute extrapolation. *Journal of Policy Analysis and Management, 33,* 527–536.

Cook, T. D. (2015). The inheritance bequeathed to William G. Cochran that he willed forward and left for others to will forward again: The limits of observational studies that seek to mimic randomized experiments. *Observational Studies, 1,* 141–164.

Cook, T. D., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). *"Sesame Street" revisited.* New York: Russell Sage Foundation.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational research* (pp. 223–326). New York: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Skokie, IL: Rand McNally.

Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin, 86,* 662–679.

Cook, T. D., Habib, F., Phillips, J., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36,* 543–597.

Cook, T. D., Scriven, M., Coryn, C. L. S., & Evergreen, S. D. H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31,* 105–117.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45,* 545–580.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27,* 724–750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, unreliable measurement and mode of data analysis. *Psychological Methods, 15,* 56–68.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability and data analysis mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research, 44,* 828–847.

Cook, T. D., Tang, Y., & Diamond, S. S. (2014). Causally valid relationships that invoke the wrong causal agent: Construct validity of the cause in policy research. *Journal of the Society for Social Work and Research, 5,* 379–414.

Cook, T. D., & Wong, V. C. (2008a). Better quasi-experimental practice. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 134–165). Thousand Oaks, CA: SAGE.

Cook, T. D., & Wong, V. C. (2008b). Empirical tests of the validity of the regression discontinuity design. *Annals of Economics and Statistics, 91/92,* 127–150.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology, 108,* 100–102.

Coryn, C. L. S., & Hobson, K. (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. In S. Mathison (Ed.), Really new directions in evaluation: Young evaluators' perspectives. *New Directions for Evaluation, 131*, 31–39.

Crespi, C. M. (2016). Improved designs for cluster randomized trials. *Annual Review of Public Health, 37,* 1–16.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. (1976). *Analysis of covariance: Angel of salvation or temptress and deluder* (Occasional papers, Stanford Evaluation Consortium). Stanford, CA: Department of Education, Stanford University.

Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. G. (1977). *Analysis of covariance in nonrandomized experiments: Parameters affecting bias* (Occasional papers, Stanford Evaluation Consortium). Stanford, CA: Department of Education, Stanford University.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *American Economic Review, 85,* 341–364.

Darley, J. M., & Latané, B. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. New York: Guilford Press.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95,* 407–424.

DeMaris, A. (2014). Combating unmeasured confounding in cross-sectional students: Evaluating instrumental-variable and Heckman selection models. *Psychological Methods, 19,* 380–397.

Denis, M. L. (1990). Assessing the validity of randomized field experiments: An example from drug abuse treatment research. *Evaluation Review, 14,* 347–373.

Denzin, N. K. (1978). *Sociological methods*. New York: McGraw-Hill.

DiNardo, J., & Lee, D. S. (2004). Economic impacts of new unionization on U.S. private sector employers: 1984–2001. *Quarterly Journal of Economics, 119,* 1383–1442.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

Dooley, D. (1995). *Social research methods* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics, 20,* 115–147.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Duncan, T. E., & Duncan, S. C. (2004a). A latent growth curve modeling approach to pooled interrupted time series analyses. *Journal of Psychopathology and Behavioral Assessment, 26,* 271–278.

Duncan, T. E., & Duncan, S. C. (2004b). An introduction to latent growth curve modeling. *Behavior Therapy, 35,* 333–363.

Eckert, W. A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation, 21,* 185–193.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology, 34,* 437–442.

Evans, R. I., Rozelle, R. M., Maxwell, S. E., Raines, B. E., Dill, C. A., Guthrie, T. J., . . . Hill, P. C. (1981). Social modeling films to deter smoking in adolescents: Results from a three-year field investigation. *Journal of Applied Psychology, 66,* 399–414.

Festinger, L. (1957). *A theory of cognitive dissonance*. Redwood City, CA: Stanford University Press.

Fetterman, D. M. (1982). Ibsen's baths: Reactivity and insensitivity. *Educational Evaluation and Policy Analysis, 4,* 261–279.

Finn, J. D., & Achilles, C. (1999). Tennessee's class size study: Findings, implications, and misconceptions. *Educational Evaluation and Policy Analysis, 21,* 97–109.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Olivier & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Flay, B. R., & Collins, L. M. (2005). Historical review of school based randomized trials for evaluation problem behavior. *Annals of the American Academy of Political and Social Science, 599,* 115–146.

Foster, M. E., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods, 1,* 249–260.

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medical Association, 303,* 47–53.

Fox, J. (2000). *Nonparametric simple regression: Smoothing scatterplots*. Thousand Oaks CA: SAGE.

Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes. *Biometrika, 86,* 365–379.

Freedman, D., Pisani, R., & Purves, R. (1978). *Statistics*. New York: Norton.

Funk, M. J., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology, 173*(7), 761–767.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8,* 172–179.

Galles, G. M. (1995, July 9). Higher speed limits may reduce traffic fatalities. *The Denver Post,* p. 1E.

Gelman, A., & Imbens, G. (2018, May 14). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*. Accepted author version posted online August 17, 2017.

Gennetian, L. A., Morris, P. A., Bos, J. M., & Bloom, H. S. (2005). Constructing instrumental variables from experimental data to explore how treatments produce effects. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 75–114). New York: Russell Sage Foundation.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2010). *Impact evaluation in practice*. Washington, DC: World Bank.

Glaser, B. G., & Strauss, A. L. (1977). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine.

Glass, G. V. (1988). Quasi-experiments: The case of interrupted time series. In R. M. Jaeger (Ed.), *Complementary methods for research in education* (pp. 445–464). Washington, DC: American Educational Research Association.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder: Colorado Associated University Press.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy, 589,* 62–93.

Goldberger, A. S. (1972a). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Paper 123-72). Madison: University of Wisconsin, Institute for Research on Poverty.

Goldberger, A. S. (1972b). *Selection bias in evaluating treatment effects: The case of interaction* (Discussion Paper 129-72). Madison: University of Wisconsin, Institute for Research on Poverty.

Goldberger, A. S. (2008). Selection bias in evaluation treatment effects: Some formal illustrations. In T. Fomby, R. C. Hill, D. L. Millimet, J. A. Smith, & E. J. Vytlacil (Eds.), *Modeling and evaluating treatment effects in economics* (pp. 1–31). Amsterdam, the Netherlands: JAI Press.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research, 35,* 472–482.

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development, 58,* 80–92.

Gollob, H. F., & Reichardt, C. S. (1991). Interpreting and estimating indirect effects assuming time lags really matter. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 243–259). Washington, DC: American Psychological Association.

Gosnell, H. F. (1927). *Getting out the vote: An experiment in the stimulation of voting.* Chicago: University of Chicago Press.

Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10,* 80–100.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Graham, J. W. (2012). *Missing data: Analysis and design*. New York: Springer.

Graham, J. W., Cumsille, P. E., & Shevock, A. (2012). Methods for handling missing data. In I. B. Weiner, J. A. Schinker, & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (2nd ed., pp. 109–141). Hoboken, NJ: Wiley.

Graham, S. E., Singer, S. D., & Willett, J. B. (2009). Modeling individual change over time. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 615–636). Thousand Oaks, CA: SAGE.

Green, D. P., Leong, T. Y., Kern, H. L., Gerber, A. S., & Larimer, C. W. (2009). Testing the accuracy of regression discontinuity analysis using experimental benchmarks. *Political Analysis, 17,* 400–417.

Greevy, R., Silber, J. H., & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics, 5,* 263–275.

Guerin, D., & MacKinnon, D. P. (1985). An assessment of the impact of the California child seat restraint requirement. *American Journal of Public Health, 75,* 142–144.

Hackman, J. R., Pearce, J. L., & Wolfe, J. C. (1978). Effects of changes in job characteristics on work attitudes and behaviors: A naturally occurring quasi-experiment. *Organizational Behavior and Human Performance, 21,* 289–304.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69,* 201–209.

Hallberg, K., Wing, C., Wong, V., & Cook, T. D. (2013). Experimental design for causal inference:

Clinical trials and regression discontinuity designs. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 1, pp. 223–236). New York: Oxford University Press.

Harding, D. J., & Seefeldt, K. S. (2013). Mixed methods and causal analysis. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 91–110). Dordrecht, the Netherlands: Springer.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12,* 247–267.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York: Guilford Press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics, 30,* 239–267.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives, 9,* 85–110.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1,* 154–169.

Hennigan, K. M., Del Rosario, M. L., Heath, L., Cook, T. D., Wharton, J. D., & Calder, B. J. (1982). Impact of the introduction of television on crime in the United States: Empirical findings and theoretical implications. *Journal of Personality and Social Psychology, 55,* 239–247.

Henry, G. T., Fortner, C. K., & Thompson, C. L. (2010). Targeted funding for educationally disadvantaged students: A regression discontinuity estimate of the impact on high school student achievement. *Educational Evaluation and Policy Analysis, 32,* 183–204.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15,* 199–236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–970.

Hollister, R. G., & Hill, J. (1995). Problems in the evaluation of community-wide initiatives. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts* (pp. 173–199). Washington, DC: Aspen Institute.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27,* 205–224.

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research* (pp. 27–51). Washington, DC: American Psychological Association.

Huck, S. W., & Sandler, H. M. (1979). *Rival hypotheses: Alternative interpretations of data based conclusions*. New York: Harper & Row.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.

Huitema, B. E. (2011). *The analysis of covariance and alternatives* (2nd ed.). New York: Wiley.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60,* 38–58.

Hyman, H. H. (1954). *Interviewing in social research*. Chicago: University of Chicago Press.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105,* 765–789.

Imbens, G. W. (2010). An economist's perspective on Shadish (2010) and West and Thoemmes (2010). *Psychological Methods, 15,* 47–55.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142,* 615–635.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York: Cambridge University Press.

Jackson, M., & Cox, D. R. (2013). The principles of experimental design and their application in sociology. *Annual Review of Sociology, 39,* 27–49.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86,* 226–244.

Jacob, R., Somers, M.-A., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. *Evaluation Review, 40,* 167–198.

Jacobs, R., Zhu, P., Somers, M.-A., & Bloom, H. (2012). *A practical guide to regression discontinuity*. Washington, DC: Manpower Demonstration Research Corporation.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Jacoby, W. G. (2000). Loess: A nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies, 19,* 577–613.

Joyce, T., Kaestner, R., & Colman, S. (2006). Changes in abortions and births and the Texas Parental Notification Law. *New England Journal of Medicine, 354,* 1031–1038.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.

Judd, C. M., Yzerbyt, V. Y., & Muller, D. (2014). Mediation and moderation. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 653–676). New York: Cambridge University Press.

Jurs, S. G., & Glass, G. V (1971). The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *Journal of Experimental Education, 40,* 62–66.

Kazdin A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford, UK: Oxford University Press.

Khuder, S. A., Milz, S., Jordan, T., Price, J., Silvestri, K., & Butler, P. (2007). The impact of a smoking ban on hospital admissions for coronary heart disease. *Preventive Medicine, 45,* 33–38.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching. Retrieved July 9, 2018, from *https://gking.harvard.edu/files/gking/files/psnot.pdf*.

Kirk, R. E. (2009). Experimental design. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 23–45). Thousand Oaks, CA: SAGE.

Kish, L. (1987). *Statistical designs for research*. New York: Wiley.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York: Guilford.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from *http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf*.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34,* 26–38.

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In R. T. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research* (pp. 91–125). Washington, DC: American Psychological Association.

Kruglanski, A. W., & Kroy, M. (1975). Outcome validity in experimental research: A reconceptualization. *Journal of Representative Research in Social Psychology, 7,* 168–178.

Lahey, B. B., & D'Onofrio, B. M. (2010). All in the family: Comparing siblings to test causal hypotheses regarding environmental influences on behavior. *Current Directions in Psychology, 19,* 319–323.

Lam, J. A., Hartwell, S. W., & Jekel, J. F. (1994). "I prayed real hard, so I know I'll get in": Living with randomization. In K. J. Conrad (Ed.), *Critically evaluating the role of experiments* (New Directions for Program Evaluation No. 63, pp. 55–66). San Francisco: Jossey-Bass.

Langer, E. J., & Rodin, J. (1976). The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. *Journal of Personality and Social Psychology, 34,* 191–198.

Lee, D. S. (2008). Randomized experiments from non-random selection in the U.S. House elections. *Journal of Econometrics, 142,* 675–697.

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics, 142,* 655–674.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48,* 281–355.

Lehman, D. R., Lampert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43,* 431–442.

Leibowitz, S. F., & Kim, T. (1992). Impact of a galanin antagonist on exogenous galanin and natural patterns of fat ingestion. *Brain Research, 599,* 148–152.

Leigh, J. P., & Schembri, M. (2004). Instrumental variables technique: Cigarette price provided better estimate of effects of smoking on SF-12. *Journal of Clinical Epidemiology, 57,* 284–293.

Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics: A rogue economist explores the hidden side of everything.* New York: Morrow.

Li, M. (2012). Using the propensity score method to estimate causal effects: A review and practical guide. *Organizational Research Methods, 16,* 188–226.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research in higher education.* Cambridge, MA: Harvard University Press.

Linden, A. (2015). Conducting interrupted time-series analysis for single- and multiple-group comparisons. *The Stata Journal, 15,* 480–500.

Lipsey, M. W., Cordray, D. S., & Berger, D. E. (1981). Evaluation of a juvenile diversion program: Using multiple lines of evidence. *Evaluation Review, 5,* 283–306.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

List, J. A., Sadoff, S., & Wagner, M. (2010). *So you want to run an experiment, now what?: Some simple rules of thumb for optimal experimental design* (Working Paper No. 15701). Cambridge, MA: National Bureau of Economic Research.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202.

Little, R. J. (1995). Modeling the drop-out mechanism in longitudinal studies. *Journal of the American Statistical Association, 90,* 1112–1121.

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health, 21,* 121–145.

Little, R. J., & Rubin, D. B. (2002). *Statistical analyses with missing data.* New York: Wiley.

Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science, 14,* 570–580.

Liu, Y., & West, S. G. (2015). Weekly cycles in daily report data: An overlooked issue. *Journal of Personality, 84,* 560–579.

Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models: An introduction to factor, path, and structural analysis*. New York: Routledge.

Lohr, B. W. (1972). *An historical view of the research on the factors related to the utilization of health services*. Rockville, MD: Social and Economic Analysis Division, Bureau for Health Services Research and Evaluation.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.

Louie, J., Rhoads, C., & Mark, J. (2016). Challenges to using the regression discontinuity design in educational evaluation: Lessons from the Transition to Algebra Study. *American Journal of Evaluation, 37,* 381–407.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29,* 530–558.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, UK: Oxford University Press.

MacKinnon, D. P., Cheong, J., & Pirlott, A. G. (2012). Statistical mediation analysis. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Vol. 2. Research designs* (pp. 313–331). Washington, DC: American Psychological Association.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49,* 301–321.

Magidson, J. (1977). Toward a causal model approach for adjusting for pre-existing differences in the nonequivalent control group situation: A general alternative to ANCOVA. *Evaluation Quarterly, 1,* 399–420.

Magidson, J., & Sörbom, D. (1982). Adjusting for confounding factors in quasi-experiments: Another reanalysis of the Westinghouse Head Start evaluation. *Educational Evaluation and Policy Analysis, 4,* 321–329.

Mandell, M. B. (2008). Having one's cake and eating it, too: Combining true experiments with regression discontinuity designs. *Evaluation Review, 32*, 415–434.

Manski, C. F., & Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociological Methodology, 28,* 99–137.

Marcantonio, R. J., & Cook, T. D. (1994). Convincing quasi-experiments: The interrupted time series and regression-discontinuity designs. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 133–154). San Francisco: Jossey-Bass.

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentation. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (New Directions for Program Evaluation No. 31, pp. 47–66). San Francisco: Jossey-Bass.

Mark, M. M., & Reichardt, C. S. (2004). Quasi-experimental and correlational designs: Methods for the real world when random assignment isn't feasible. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *Handbook of methods in social psychology* (pp. 265–286). Thousand Oaks, CA: SAGE.

Mark, M. M., Reichardt, C. S., & Sanna, L. J. (2000). Time-series designs and analyses. In H. E. A. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 353–389). New York: Academic Press.

Marsh, J. C. (1985). Obstacles and opportunities in the use of research on rape legislation. In R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 295–310). Beverly Hills, CA: SAGE.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavioral Research, 44,* 1–23.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12,* 23–44.

Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research, 46,* 816–841.

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin, 95,* 136–147.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Mahwah, NJ: Erlbaum.

May, H. (2012). Nonequivalent comparison group designs. In H. Cooper (Ed.), *The APA handbook of research methods in psychology* (pp. 489–509). Washington, DC: American Psychological Association.

Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total and indirect causal effects. *Multivariate Behavioral Research, 49,* 425–442.

Mazza, G. L., & Enders, C. (2014). Missing data analysis. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 627–652). New York: Cambridge University Press.

McCleary, R., McDowall, D., & Bartos, B. (2017). *Design and analysis of time series experiments*. New York: Oxford University Press.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142,* 698–714.

McGinley, L. (1997, October 29). Saccharin's presence on list of carcinogens may be near an end. *The Wall Street Journal,* pp. A1, A10.

McKillip, J. (1992). Research without control groups: A control construct design. In R. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, & E. Henderson (Eds.), *Methodological issues in applied psychology* (pp. 159–175). New York: Plenum Press.

McLeod, R. S., Taylor, D. W., Cohen, A., & Cullen, J. B. (1986). Single patient randomized clinical trials: Its use in determining optimal treatment for patient with inflammation of a Kock continent ileostomy reservoir. *Lancet, 327,* 728–729.

McSweeny, A. J. (1978). Effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavior Analysis, 11,* 47–51.

Mees, C. E. K. (1934). Scientific thought and social reconstruction. *Sigma Xi Quarterly, 22,* 13–24.

Meier, P. (1972). The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters,

& G. R. Rising (Eds.), *Statistics: A guide to the unknown* (pp. 120–129). San Francisco: Holden-Day.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14,* 5–8.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics, 13,* 151–161.

Mills, J. L. (1993). Data torturing. *New England Journal of Medicine, 329,* 1196–1199.

Millsap, M. A., Goodson, B., Chase, A., & Gamse, B. (1997). *Evaluation of "Spreading the Comer School Development Program and Philosophy."* Cambridge, MA: Abt.

Minton, J. H. (1975). The impact of Sesame Street on reading readiness of kindergarten children. *Sociology of Education, 48,* 141–151.

Mohler, D., Schultz, K. F., & Altman, D. G., for the CONSORT Group. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet, 357,* 1191–1194.

Mohr, L. B. (1995). *Impact analysis for program evaluation* (2nd ed.). Thousand Oaks, CA: SAGE.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halback, A., & Ehrle, K. (1999). Evaluating the SAGE Program: A pilot program in targeted pupil–teacher reduction in Wisconsin. *Educational Evaluation and Policy Analysis, 21,* 165–177.

Mosca, J. B., & Howard, L. W. (1997). Grounded learning: Breathing life into business education. *Journal of Education for Business, 73,* 90–93.

Moskowitz, J. M. (1993). Why reports of outcome evaluations are often biased or uninterpretable: Examples from evaluations of drug abuse prevention programs. *Evaluation and Program Planning, 16,* 1–9.

Moss, B. G., Yeaton, W. H., & Lloyd, J. E. (2014). Evaluating the effectiveness of developmental mathematics by embedding a randomized experiment within a regression discontinuity design. *Educational Evaluation and Policy Analysis, 36,* 170–185.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research.* New York: Oxford University Press.

Murray, D. M. (1998). *Design and analysis of group randomized trials.* New York: Oxford University Press.

Murray, D. M., Moskowitz, J. M., & Dent, C. W. (1996). Design and analysis issues in community-based drug abuse prevention. *American Behavioral Scientist, 39,* 853–867.

Nugent, W. R. (2010). *Analyzing single system design data.* Oxford, UK: Oxford University Press.

Nugent, W. R., Champlin, D., & Wiinimaki, L. (1997). The effects of anger control training on adolescent antisocial behavior. *Research on Social Work Practice, 7*(4), 446–462.

Orne, M. T. (1963). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 358–372.

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in behavioral research* (pp. 143–179). New York: Academic Press.

Orr, L. L., Bloom, H. S., Bell, S. H., Doolittle, F., & Lin, W. (1996). *Does training for the disadvantaged work?: Evidence from the National JTPA study.* Washington: DC: Urban Institute Press.

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics, 161,* 203–207.

Patry, J.-L. (2013). Beyond multiple methods: Critical multiplism on all levels. *International Journal of Multiple Research Approaches, 7,* 50–65.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill & Wang.
Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity score matching: A note of caution for evaluators of social programs. *The American Statistician, 62,* 222–231.
Pennel, M. L., Hade, E. M., Murray, D. M., & Rhoda, D. A. (2011). Cutoff designs for community-based intervention studies. *Statistics in Medicine, 30,* 1865–1882.
Pischke, J.-S. (2007). The impact of length of the school year on student performance and earnings: Evidence from the German short school year. *Economic Journal, 117,* 1216–1242.
Pitts, S. C., West, S. G., & Tein, J.-Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning, 19,* 333–350.
Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis, 31,* 463–479.
Porter, A. C. (1967). The effects of using fallible variables in the analysis of covariance. (Doctoral dissertation, University of Wisconsin). Dissertation Abstracts International, 1968, 28, 3517B. (University Microfilms No. 67–12, 147, 144.)
Porter, A. C., & Chibucos, T. R. (1974). Selecting analysis strategies. In G. D. Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, NJ: Educational Technology Publication.
Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research, 15,* 213–234.
Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology, 66,* 825–852.
Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185.
Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*(4), 25–31.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.
Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.
Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluations and Policy Analysis, 29,* 5–29.
Reding, G. R., & Raphelson, M. (1995). Around-the-clock mobile psychiatric crisis intervention: Another effective alternative to psychiatric hospitalization. *Community Mental Health Journal, 31,* 179–187.
Redmond, T. J. (2016). Political obfuscation. *Skeptic, 21,* 54–59.
Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell, *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–205). Chicago: Rand McNally.
Reichardt, C. S. (2000). A typology of strategies for ruling out threats to validity. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 89–115). Thousand Oaks, CA: SAGE.
Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods, 11,* 1–18.
Reichardt, C. S. (2009). Quasi-experimental design. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 46–71). Thousand Oaks, CA: SAGE.
Reichardt, C. S. (2011a). Commentary: Are three waves of data sufficient for assessing mediation? *Multivariate Behavioral Research, 46,* 842–851.

Reichardt, C. S. (2011b). Criticisms of and an alternative to the Shadish, Cook and Campbell Validity Typology. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing validity in outcome evaluation: Theory and practice* (New Directions for Evaluation No. 130, pp. 43–53). Hoboken, NJ: Wiley.

Reichardt, C. S. (2011c). Evaluating methods for estimating program effects. *American Journal of Evaluation, 32,* 246–272.

Reichardt, C. S., & Gollob, H. F. (1986). Satisfying the constraints of causal modeling. In W. M. K. Trochim (Ed.), *Advances in quasiexperimental design and analysis* (New Directions for Program Evaluation, No. 31, pp. 91–107). San Francisco: Jossey-Bass.

Reichardt, C. S., & Gollob, H. F. (1987). Taking uncertainty into account when estimating effects. In M. M. Mark & R. L. Shotland (Eds.), *Multiple methods in program evaluation* (New Directions for Program Evaluation No. 35, pp. 7–22). San Francisco: Jossey-Bass.

Reichardt, C. S., & Gollob, H. F. (1989). Ruling out threats to validity. *Evaluation Review, 13,* 3–17.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259–284). Hillsdale, NJ: Erlbaum.

Reichardt, C. S., & Gollob, H. F. (1999). Justifying the use and increasing the power of a *t* test for a randomized experiment with a convenience sample. *Psychological Methods, 4,* 117–128.

Reichardt, C. S., & Henry, G. T. (2012). Regression-discontinuity designs. In H. Cooper (Ed.), *The APA handbook of research methods in psychology* (pp. 511–526). Washington, DC: American Psychological Association.

Reichardt, C. S., & Mark, M. M. (1998). Quasi-experimentation. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 193–228). Thousand Oaks, CA: SAGE.

Reichardt, C. S., & Rallis, S. F. (1994). The relationship between the qualitative and quantitative research traditions. In C. S. Reichardt & S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives* (New Directions for Program Evaluation No. 61, pp. 5–11). San Francisco: Jossey-Bass.

Reichardt, C. S., & Trochim, W. M. K., & Cappelleri, J. C. (1995). Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review, 19,* 39–63.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11,* 691–714.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19,* 1–25.

Riecken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. K., Jr., Pratt, J. W., . . . Williams, W. (1974). *Social experimentation: A method for planning and evaluating social intervention*. New York: Academic Press.

Rindskopf, D. M., & Ferron, J. M. (2014). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 221–246). Washington, DC: American Psychological Association.

Roos, L. L., Jr., Roos, N. P., & Henteleff, P. D. (1978). Assessing the impact of tonsillectomies. *Medical Care, 16,* 502–518.

Rosenbaum, P. R. (1984a). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P. R. (1984b). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society,* Series A, *147,* 656–666.
Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11,* 207–224.
Rosenbaum, P. R. (1987). The role of a second control group in an observational study. *Statistical Science, 2,* 292–316.
Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115,* 901–905.
Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science, 14,* 259–304.
Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.
Rosenbaum, P. R. (2005a). Observational study. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1451–1462). New York: Wiley.
Rosenbaum, P. R. (2005b). Sensitivity analysis in observational studies. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1809–1814). New York: Wiley.
Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.
Rosenbaum, P. R. (2013). Using differential comparisons in observational studies. *Chance, 26,* 18–25.
Rosenbaum, P. R. (2015a). Cochran's causal crossword. *Observational Studies, 1,* 205–211.
Rosenbaum, P. R. (2015b). How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and Its Applications, 2,* 21–48.
Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Cambridge, MA: Harvard University Press.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.
Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational students using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.
Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
Rosenthal, R., & Rosnow, R. L. (Eds.). (1969). *Artifact in behavioral research*. New York: Academic Press.
Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York: Freeman.
Ross, H. L., Campbell, D. T., & Glass, G. V (1970). Determining the social effects of a legal reform: The British "Breathalyser" crackdown of 1967. *The American Behavioral Scientist, 13,* 493–509.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.
Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26.
Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34–58.
Rubin, D. B. (1986). Comment: Which ifs have causal answers?: Discussion of Holland's "Statistics and causal inference." *Journal of the American Statistical Association, 81,* 961–962.
Rubin, D. B. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association, 95,* 435–438.
Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics, 29,* 343–367.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26,* 20–36.

Rubin, D. B. (2008a). Objective causal inference, design trumps analysis. *Annals of Applied Statistics, 2,* 808–840.

Rubin, D. B. (2008b). The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association, 103,* 1350–1356.

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods, 15,* 38–46.

Russell, B. (1913). On the notion of cause. *Proceedings of the Aristotelian Society, 13,* 1–26.

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., & Ritchie, T. D. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods, 19,* 317–333.

Schachter, S. (1982). Recidivism and self-cure of smoking and obesity. *American Psychologist, 37,* 436–444.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13,* 279–313.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (Report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Schochet, P. Z. (2008). *Technical Methods Report: Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics, 34,* 238–266.

Schoemann, A. M., Rhemtulla, M., & Little, T. (2014). Multilevel and longitudinal modeling. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 571–588). New York: Cambridge University Press.

Scriven, M. (1968). In defense of all causes. *Issues in Criminology, 4,* 79–81.

Scriven, M. (1975). Causation as explanation. *Nous, 9,* 3–16.

Scriven, M. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 101–118). Beverly Hills, CA: SAGE.

Scriven, M. (2008). A summative evaluation of RCT methodology: And an alternative approach to causal research. *Journal of Multidisciplinary Evaluation, 5,* 11–24.

Scriven, M. (2009). Demythologizing causation and evidence. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 134–152). Thousand Oaks, CA: SAGE.

Seaver, W. B., & Quarton, R. J. (1976). Regression-discontinuity analysis of dean's list effects. *Journal of Educational Psychology, 66,* 459–465.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of research. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15–25). Beverly Hills, CA: SAGE.

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine, 13,* 1715–1726.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. In L. B. Sechrest & A. J. Figueredo (Eds.), *Program evaluation: A pluralistic enterprise* (New directions for Program Evaluation No. 60, pp. 13–57). San Francisco: Jossey-Bass.

Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 13–25). Thousand Oaks, CA: SAGE.

Shadish, W. R. (2002). Revisiting field experiments: Field notes for the future. *Psychological Methods, 7,* 3–18.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15,* 3–17.

Shadish, W. R. (2013). Propensity score analysis: Promise, reality, and irrational exuberance. *Journal of Experimental Criminology, 9,* 129–144.

Shadish, W. R. (2014a). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52,* 109–122.

Shadish, W. R. (2014b). Statistical analysis of single-case designs: The shape of things to come. *Current Directions in Psychological Science, 23,* 139–146.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers?: A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103,* 1334–1343.

Shadish, W. R., & Cook, T. D. (1999). Comment—design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science, 14,* 294–300.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology, 60,* 607–629.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Shadish, W. R., Cook, T. D., & Houts, A. C. (1986). Quasi-experimentation in a critical multiplist mode. In W. M. K. Trochim, *Research design for program evaluation: The regression-discontinuity approach* (pp. 29–46). Beverly Hills, CA: SAGE.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random to cut-off based assignment. *Psychological Methods, 16,* 179–191.

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Rindskopf, D. M., Boyajian, J. G., & Sullivan, K. J. (2014). Analyzing single-case designs: d, G, hierarchical models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analyses. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 247–281). Washington, DC: American Psychological Association.

Shadish, W. R., Hu, X., Glaser, R. R., Knonacki, R. J., & Wong, T. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3,* 3–22.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18,* 385–405.

Shadish, W. R., & Luellen, J. K. (2006). Quasi-experimental design. In J. Green, G. Camilli, & P. Elmore (Eds.), *Complementary methods for research in education* (pp. 539–550). Mahwah, NJ: Erlbaum.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin, 126,* 512–529.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazelrigg, M. D.,

. . . Weiss, B. (1997). Evidence that therapy works in clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65,* 355–365.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in psychotherapy experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64,* 1290–1305.

Shadish, W. R., & Sullivan, K. J. (2012). Theories of causation in psychological science. In H. Cooper (Ed.), *The APA handbook of research methods in psychology* (pp. 23–52). Washington, DC: American Psychological Association.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology, 52,* 41–70.

Sheiner, L. B., & Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology and Therapy, 57,* 6–10.

Sheridan, S. M. (2014). Single-case designs and large-N studies: The best of both worlds. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research* (pp. 299–308). Washington, DC: American Psychological Association.

Sherman, L. W., & Weisburd, D. (1995). General deterrent effects of policy patrol in crime hotspots: A randomized controlled trial. *Justice Quarterly, 12,* 625–648.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6,* 317–329.

Smith, H. F. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics, 13,* 282–308.

Smith, H. L. (2013). Research design: Toward a realistic role for causal analysis. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 45–73). Dordrecht, the Netherlands: Springer.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17,* 510–550.

Smith, M. L., Gabriel, R., Schoot, J., & Padia, W. L. (1976). Evaluation of the effects of Outward Bound. In G. V Glass (Ed.), *Evaluation Studies Review Annual* (Vol. 1, pp. 400–421). Newbury Park, CA: SAGE.

Snow, J. (1855). *On the mode of communication of cholera* (2nd ed.). London: John Churchill.

Snyder, S. H. (1974). *Madness and the brain*. New York: McGraw-Hill.

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. Retrieved from *www.mdrc.org/publication/validity-and-precision-comparative-interrupted-time-series-design-and-difference*.

Sovey, A. J., & Green, D. P. (2011). Instrumental variables estimation in political science: A reader's guide. *American Journal of Political Science, 55,* 188–200.

Sparks, S. D. (2010, October 20). "What Works" broadens its research standards: Clearinghouse moves past "gold standard." Retrieved from *www.edweek.org/ew/articles/2010/10/20/08wwc.h30.html*.

Splawa-Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments: Essay on principles, Section 9 (D. M. Dabrowska & T. P. Speed, Trans.). *Statistical Science, 5,* 465–472.

Sprinthall, R. C. (1997). *Basic statistical analysis* (5th ed.). Boston: Allyn & Bacon.

St. Clair, T., & Cook, T. D. (2015). Differences-in-differences methods in public finance. *National Tax Journal, 2,* 319–338.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation, 35,* 311–327.

St. Clair, T., Hallberg, K., & Cook, T. D. (2016). The validity and precision of the comparative interrupted time-series design. *Journal of Educational and Behavioral Statistics, 41,* 269–299.

St. Pierre, R. G., Ricciuti, A., & Creps, C. (2000). *Synthesis of local and state Even Start evaluations*. Cambridge, MA: Abt.

Steiner, D., & Mark, M. M. (1985). The impact of a community action group: An illustration of the potential of time series analysis for the study of community groups. *American Journal of Community Psychology, 13,* 13–30.

Steiner, P. M., & Cook, D. (2015). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 1, pp. 237–259). New York: Oxford University Press.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustment using propensity scores. *Journal of Educational and Behavioral Statistics, 36,* 213–236.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15,* 250–267.

Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review, 62,* 494–507.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155–176). Thousand Oaks, CA: SAGE.

Sullivan, C. M., Rumptz, M. H., Campbell, R., Eby, K. K., & Davidson, W. S. (1996). Retaining participants in longitudinal community research: A comprehensive protocol. *Journal of Applied Behavioral Science, 32,* 262–276.

Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2015). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological Methods, 20,* 26–42.

Tang, Y., Cook, T. D., & Kisbu-Sakarya, Y. (2018). Statistical power for the comparative regression discontinuity design with a nonequivalent comparison group. *Psychological Methods, 23,* 150–168.

Ten Have, T. R., Normand, S. T., Marcus, S. M., Brown, C. H., Lavori, P., & Duan, N. (2008). Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatric Annals, 38,* 772–783.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post-facto experiment. *Journal of Educational Psychology, 51,* 309–317.

Thoemmes, F., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46,* 90–118.

Thoemmes, F., Liao, W., & Jin, Z. (2017). The analysis of the regression-discontinuity design in R. *Journal of Educational and Behavioral Statistics, 42,* 341–360.

Travers, R. M. W. (1981, June/July). Letter to the Editor. *Educational Researcher,* p. 32.

Trochim, W. M. K. (1984). *Research designs for program evaluation: The regression-discontinuity approach*. Newbury Park, CA: SAGE.

Trochim, W. M. K. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9,* 575–604.

Trochim, W. M. K. (1989). Outcome pattern matching and program theory. *Evaluation and Program Planning, 12,* 355–366.

Trochim, W. M. K., & Cappelleri, J. C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials, 13,* 190–212.

Trochim, W. M. K., Cappelleri, J. C., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review, 15,* 571–604.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review, 99,* 179–215.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (2017). Standards handbook: Version 4.0. Retrieved from *https://ies.ed.gov/ncee/wwc/handbooks*.

U.S. General Accounting Office. (1991, July). *Motorcycle helmet laws save lives and reduce costs to society* (GAO/RCED-91-170). Washington, DC: Author.

van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression discontinuity approach. *International Economic Review, 43,* 1249–1287.

van der Klaauw, W. (2008). Regression discontinuity analysis: A survey of recent developments in economics. *LABOUR, 22,* 219–245.

van Helmont, J. B. (1662). *Oriatrik or, physick refined: The common errors therein refuted and the whole art reformed and rectified*. London: Lodowick-Lloyd.

Velicer, W. F., & Harrop, J. W. (1983). The reliability and accuracy of time series model identification. *Evaluation Review, 7*(4), 551–560.

Velicer, W. F., & McDonald, R. P. (1984). Time series analysis without model identification. *Multivariate Behavioral Research, 19,* 33–47.

W. K. Kellogg Foundation. (2004). *Logic model development guide*. Battle Creek, MI: Author.

Wagenaar, A. C. (1981). Effects of the raised legal drinking age on motor vehicle accidents in Michigan. *HSRI Research Review, 11,* 1–8.

Wagenaar, A. C. (1986). Preventing highway crashes by raising the legal minimum age for drinking: The Michigan experience 6 years later. *Journal of Safety Research, 17,* 101–109.

Wallach, M. A., & Kogan, N. (1965). The roles of information, discussion, and consensus in group risk taking. *Journal of Experimental Social Psychology, 1,* 1–19.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70,* 129–133.

Webb, E. J., & Ellsworth, P. C. (1975). On nature and knowing. In H. W. Sinaiko & L. A. Broedling (Eds.), *Perspectives on attitude measurement: Surveys and their alternatives*. Washington, DC: Smithsonian Institution.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578–1585.

West, S. G., Aiken, L. S., & Todd, M. (1993). Probing the effects of individual components in multiple component prevention programs. *American Journal of Community Psychology, 21,* 571–605.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). New York: Cambridge University Press.

West, S. G., Cham, H., & Liu, Y. (2014). Causal inference and generalizations in field settings:

Experimental and quasi-experimental designs. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 49–80). New York: Cambridge University Press.

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and applications in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology, 82,* 906–919.

West, S. G., & Graziano, W. G. (2012). Basic, applied, and full-cycle social psychology: Enhancing causal generalization and impact. In D. T. Kenrick, N. J. Goldstein, & S. L. Braver, *Six degrees of social influence: Science, application, and the psychology of Bob Cialdini* (pp. 119–133). New York: Oxford University Press.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality, 59,* 609–662.

West, S. G., Hepworth, J. T., McCall, M. A., & Reich, J. W. (1989). An evaluation of Arizona's July 1982 drunk driving law: Effects on the city of Phoenix. *Journal of Applied Social Psychology, 19,* 1212–1237.

West, S. G., & Sagarin, B. (2000). Participant selection and loss in randomized experiments. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 117–154). Thousand Oaks, CA: SAGE.

West, S. G., & Thoemmes, F. (2008). Equating groups. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 414–430). Thousand Oaks, CA: SAGE.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15,* 18–37.

Widaman, K. F. (2006). III. Missing data: What to do with or without them. In K. McCartney, M. R. Burchinal, & K. L. Bub (Eds.), *Best practices in quantitative methods for developmentalists* (pp. 42–64). *Monographs of the Society for Research in Child Development, 71* (3, Serial No. 285).

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management, 32,* 853–877.

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices in public health policy research. *Annual Review of Public Health, 39,* 453–469.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 659–707.

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness, 8,* 245–279.

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics, 38,* 107–141.

Wong, V. C., & Wing, C. (2016). The regression discontinuity design and the social corruption of quantitative indicators. *Observational Studies, 2,* 183–209.

Wong, V. C., Wing, C., Steiner, P. M., Wong, N., & Cook, T. D. (2012). Research designs for program evaluation. In J. A. Schinker & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (2nd ed., pp. 316–341). Hoboken, NJ: Wiley.

Wortman, P. M., Reichardt, C. S., & St. Pierre, R. G. (1978). The first year of the educational voucher demonstration: A secondary analysis of student achievement test scores. *Evaluation Quarterly, 2,* 193–214.

Wu, W., West, S. G., & Hughes, J. N. (2008). Effect of retention in first grade on children's achievement trajectories over 4 years. *Journal of Educational Psychology, 100*, 727–740.

Yang, M., & Maxwell, S. E. (2014). Treatment effects in randomized longitudinal trials with different types of nonignorable dropout. *Psychological Methods, 19,* 188–210.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156–167.

Yin, R. K. (2008). *Third update of student achievement data and findings, as reported in MSPs' annual and evaluators' reports.* Washington, DC: Cosmos Corporation.

Zajonc, R. B., & Markus, H. (1975). Birth order and intellectual development. *Psychological Review, 82,* 74–88.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation, 30,* 44–61.

Author Index

Subject Index

Note. *f* or *t* following a page number indicates a figure or table

About the Author

Charles S. Reichardt, PhD, is Professor of Psychology at the University of Denver. He is an elected Fellow of the American Psychological Society, an elected member of the Society of Multivariate Experimental Psychology, and a recipient of the Robert Perloff President's Prize from the Evaluation Research Society and the Jeffrey S. Tanaka Award from the Society of Multivariate Experimental Psychology. Dr. Reichardt's research focuses on quasi-experimentation.