THE PHILOSOPHY OF QUANTITATIVE METHODS

UNDERSTANDING STATISTICS

BRIAN D. HAIG

OXFORD

THE PHILOSOPHY OF QUANTITATIVE METHODS

SERIES IN UNDERSTANDING STATISTICS

S. NATASHA BERETVAS Series Editor

SERIES IN UNDERSTANDING MEASUREMENT

S. NATASHA BERETVAS Series Editor

SERIES IN UNDERSTANDING QUALITATIVE RESEARCH

PATRICIA LEAVY Series Editor

Understanding Statistics

Exploratory Factor Analysis
Leandre R. Fabrigar and Duane T. Wegener

The Philosophy of Quantitative Methods
Brian D. Haig

Validity and Validation
Catherine S. Taylor

Understanding Measurement

Item Response Theory
Christine DeMars

Reliability
Patrick Meyer

Understanding Qualitative Research

Autoethnography
Tony E. Adams, Stacy Holman Jones, and Carolyn Ellis

Qualitative Interviewing
Svend Brinkmann

Evaluating Qualitative Research: Concepts, Practices, and Ongoing Debates
Jeasik Cho

Video as Method
Anne M. Harris

Focus Group Discussions
Monique M. Hennink

The Internet
Christine Hine

Diary Methods
Lauri L. Hyers

Oral History
Patricia Leavy

Using Think-Aloud Interviews and Cognitive Labs in Educational Research
Jacqueline P. Leighton

Qualitative Disaster Research
Brenda D. Phillips

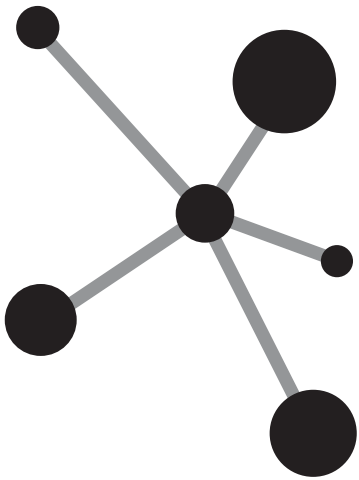
Fundamentals of Qualitative Research
Johnny Saldaña

Duoethnography
Richard D. Sawyer and Joe Norris

Analysis of the Cognitive Interview in Questionnaire Design
Gordon B. Willis

BRIAN D. HAIG

THE PHILOSOPHY OF QUANTITATIVE METHODS



OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

CIP data is on file at the Library of Congress
ISBN 978–0–19–022205–5

9 8 7 6 5 4 3 2 1

Printed by WebCom, Inc., Canada

CONTENTS

ACKNOWLEDGMENTS

In preparing this book, I have made use of previously published material from the following sources:

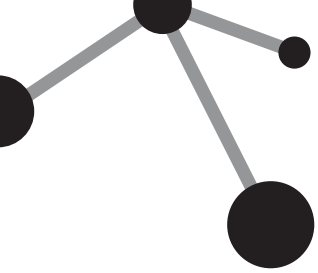
Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, *40*, 303–329.

Haig, B. D. (2012). The philosophy of quantitative methods. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (Vol. 1, pp. 6–30). New York, NY: Oxford University Press.

Haig, B. D. (2016). Tests of statistical significance made sound. *Educational and Psychological Measurement*, *77*, 489–506.

I am grateful to the journals and publishers for allowing me to make use of this material.

THE PHILOSOPHY OF QUANTITATIVE METHODS



1

INTRODUCTION

THE PHILOSOPHY of research methods is an area of knowledge that receives limited attention in behavioral research methodology and science education. The majority of students and research practitioners in the behavioral sciences obtain the bulk of their knowledge of research methods from textbooks. However, a casual examination of these texts shows that they tend to pay little, if any, serious regard to the philosophy of science and its bearing on the research process. As Thomas Kuhn pointed out more than 50 years ago (Kuhn, 1962/1996), textbooks play a major role in dogmatically initiating students into the routine practices of normal science. Serious attention to the philosophy of research methods would go a considerable way toward overcoming this uncritical practice.

This book is concerned with the philosophical foundations of research methods. In particular, it undertakes a philosophical examination of a number of different quantitative research methods that are prominent in, or relevant for, the conduct of research in the behavioral sciences. The methods submitted to critical examination are exploratory data analysis, statistical significance testing, Bayesian confirmation theory and statistics, meta-analysis, and exploratory factor analysis. I introduce these methods, and explain their selection, in the overview section that follows.

The Philosophy of Quantitative Methods

Historically, philosophers of science have given research methods in science limited attention, concentrating mostly on the nature and purpose of theory in the physical sciences. More recently, however, they have shown an increased willingness to deal with methodological issues in sciences other than physics, particularly biology, but also psychology and related behavioral and social sciences to some extent. In short, there is a developing literature in contemporary philosophy of science that can aid both our understanding, and use, of a variety of research methods and strategies in psychology. Increasingly, the philosophy of science contributes important methodological insights that are impossible to ignore when coming to grips with research methods. Increasingly, the philosophy of science is becoming a philosophy *for* science. At the same time, a miscellany of theoretically oriented psychologists, and behavioral and social scientists more generally, has produced work on the conceptual foundations of research methods that helps illuminate those methods. The work of both professional philosophers of science and theoretical scientists deserves to be included in a philosophical examination of behavioral research methods.

The three major philosophies of science that bear on psychology are empiricism, social constructionism, and scientific realism (Greenwood, 1992; Manicas & Secord, 1983). Nineteenth-century British empiricism had a major influence on the development of British statistics in the first half of the twentieth century (Mulaik, 1985). The statistical methods developed in that intellectual milieu remain an important part of psychology's statistical research practice. For example, Karl Pearson's product moment correlation coefficient was taken by its founder to be the quantitative expression of a causal relation viewed in empiricist terms. Similarly, Fisher's endorsement of inductive methods as the proper view of scientific method stemmed from a commitment to the empiricism of his day. Even in the current postpositivist philosophical climate, authors of research methods textbooks sometimes portray quantitative research as essentially positivist in its empiricist commitments (Yu, 2006). Among other things, positivism restricts its attention to what can be observed and regards theories as instruments that organize claims about observables but do not explain them by appeal to hidden causes.

Qualitative methodologists also often bolster their preferred conception of qualitative research by comparing it with an unflattering positivist picture of quantitative research. They tend to adopt the philosophy of social constructionism, which is opposed to the traditional notions of truth, objectivity, and reason, maintaining that our understanding of the world is determined by social negotiation. In one or other of its various forms, it is the philosophy of choice for many qualitative researchers, and it tends to be employed by those who are opposed, or indifferent, to quantitative methods.

Scientific Realism and Its Methodology

The book adopts a scientific realist perspective on research methods, although its emphasis is not always evident. Scientific realism, like all philosophies of science, is the subject of considerable debate, and it is opposed by many antirealist positions (principally, the philosophies of empiricism and social constructivism). Nonetheless, with justification, it remains the dominant philosophy of science to this day (Psillos, 1999). It is also the tacit philosophy of most working scientists. This fact, combined with its current emphasis on the nature of scientific practice, makes scientific realism the philosophy of choice *for* science.

Scientific realism comes in many forms. Most versions of scientific realism display a commitment to at least two doctrines: (1) that there is a real world of which we are part and (2) that both the observable and unobservable features of that world can be known by the proper use of scientific methods. Some versions of scientific realism incorporate additional theses (e.g., the claims that truth is the primary aim of science, and that successive theories more closely approximate the truth), and some will also nominate optional doctrines that may, but need not, be used by scientific realists (e.g., the claim that causal relations are relations of natural necessity; see Hooker, 1987). Others who opt for an "industrial strength" version of scientific realism for the physical sciences are more cautious about its successful reach in the behavioral sciences. In philosophy, J. D. Trout (1998), for example, subscribes to a modest realism in psychology, based on his skepticism about the discipline's ability to produce deeply informative theories like those of the physical sciences. In psychology, James Grice (2011) presents a

philosophy of moderate realism to underwrite his novel methodology of "observation oriented modeling." This philosophy maintains that things have essences, that their natures are knowable, and that a strategy of modeling can be used to integrate knowledge about the systems under study. Grice shows in a general way how philosophy of science can make an important contribution to scientific methodology.

Scientific realism boasts a rich conception of methodology, which is of considerable help in understanding and guiding research. The resourcefulness of realist methodology is suggested in the following description of its major characteristics (cf. Haig, 2014; Hooker, 1987; Nickles, 1987): First, realist methodology has three major tasks: to describe how methods function; to evaluate methods critically against their rivals; and to recommend how to use particular methods to pursue chosen research goals. I hope that my concern with these tasks is evident in the treatment of the methods in the following chapters.

Second, realist methodology is critically aim oriented. At a broad level, it recommends the pursuit of valuable truth, explanatory understanding, and effective control as primary research goals; it is also concerned with the mutual adjustment of methods and research goals. At a more specific level, my discussion of methods attempts to give due recognition to their appropriate research goals.

Third, realist methodology is naturalistic; that is to say, it is a substantive domain that uses the methods of the various sciences to study methods themselves. The error-statistical perspective presented in Chapter 3 is a philosophy of statistics that sits squarely within the naturalistic tradition in modern philosophy. Proctor and Capaldi (2001) advocate a naturalistic approach to methodology in psychology in which the empirical justification of methodological ideas is emphasized.

A fourth feature of realist methodology is that it is both generative and consequentialist. Generative methodology involves reasoning to, and accepting, knowledge claims in question from warranted premises. Exploratory factor analysis is a prominent example of a method in psychology that involves a generative justification of the factorial hypotheses to which it gives rise. By contrast, consequentialist methodology focuses on reasoning from knowledge claims in question to their testable consequences. The

widely used hypothetico-deductive method, with its emphasis on predictive accuracy, clearly exhibits a consequentialist approach to justifying knowledge claims.

Fifth, realist methodology acknowledges the need for two quite different approaches to justifying knowledge claims. In philosophy, these are commonly known as *reliabilism* and *coherentism*. With reliabilism, a belief is justified to the extent that it is acquired by reliable processes. In general, the innumerable methods that contribute to the detection of empirical phenomena are concerned with reliabilist justification. With coherentism, a belief is justified in virtue of its coherence with other beliefs. Thagard's (1992) theory of explanatory coherence (which is not considered in this book) is used for the comparative evaluation of scientific theories and embodies an illuminating coherentist perspective on knowledge justification. These two forms of justification are different, complementary, and of equal importance.

As a sixth feature, realist methodology regards science as a problem-oriented endeavor in which problems are conceptualized as constraints on their effective solution (Haig, 1987; Nickles, 1981). On this formulation, the constraints are actually constitutive of the problem itself; they characterize the problem and give it structure. Further, by including all the constraints in the problem's articulation, the problem enables the researcher to direct inquiry effectively by pointing the way to its own solution. In a real sense, stating the problem is half the solution! This focus on research problems holds more promise for research inquiry than the customary talk about research questions.

Finally, realist methodology takes the researcher's makeup as a "knowing subject" seriously. Among other things, the researcher is regarded as a satisficer who makes heavy use of heuristics to guide inquiries. McGuire (1997), for example, discusses many useful heuristics that can be employed to facilitate the generation of hypotheses in psychological research.

Scientific realist methodology undergirds a wide variety of methods, strategies, and heuristics that have been successfully used to produce worthwhile knowledge about both empirical phenomena and explanatory theories. If quantitative researchers in psychology fully engage this literature, they will find resources for enhancing their understanding of research methods and the proper uses to which they can be put.

Theories of Scientific Method

Modern science is a multifaceted endeavor. A full appreciation of its nature needs to consider the aims it pursues, the theories it produces, the methods it employs, and the institutions in which it is embedded. Although all of these features are integral to science, science is most illuminatingly characterized as method. Method is central to science because much of what we have learned from science has been acquired through use of its methods. Our scientific methods have been acquired in the course of learning about the world; as we learn, we use methods and theorize about them with increased understanding and success. Applied to science, *method* suggests the efficient, systematic ordering of inquiry. Scientific method, then, describes a sequence of actions that constitute a strategy to achieve one or more research goals. Relatedly, scientific *methodology* denotes the general study of scientific methods and forms the basis for a proper understanding of those methods. Modern scientific methodology has given considerable attention to a number of general theories of scientific method. Here, I sketch three theories that figure in the chapters that follow: inductive method, hypothetico-deductive method, and abductive method. These theories of method provide different orientations to the more specific research methods considered in Chapters 2–6.

The idea that scientific method involves inductive reasoning goes back at least to Aristotle and was given heavy emphasis by Francis Bacon and John Stuart Mill. Inductive reasoning takes different forms. For example, it is to be found in the fashioning of statistical generalizations, in the Bayesian assignment of probabilities to hypotheses, and even in the reasoning involved in moving from data to hypotheses in the hypothetico-deductive method. In psychology, the radical behaviorism of B. F. Skinner is a prominent example of a research tradition that makes use of an inductive conception of scientific method.

The most popular account of method in science is the hypothetico-deductive method. It has come to assume hegemonic status in the behavioral sciences and places a heavy emphasis on testing hypotheses in terms of their predictive success. Relatedly, the use of traditional statistical significance test procedures in psychology is often embedded in a hypothetico-deductive structure. With the hypothetico-deductive method, the scientist takes

a hypothesis or a theory and tests it indirectly by deriving from it one or more observational predictions that are amenable to direct empirical test. If the predictions are borne out by the data, then that result is taken as a confirming instance of the theory in question. If the predictions fail to square with the data, then that fact counts as a disconfirming instance of the theory.

According to the abductive theory of method (Haig, 2014), scientific inquiry is a problem-solving endeavor in which sets of data are analyzed to detect robust empirical regularities, or phenomena. Once detected, these phenomena are explained by abductively inferring the existence of underlying causal mechanisms. On positive judgments of the initial plausibility of these explanatory theories, attempts are made to elaborate on the nature of the causal mechanisms in question. This is done by constructing plausible models of those mechanisms by analogy with relevant ideas in domains that are well understood. When the theories are well developed, they are assessed against their rivals in respect of their explanatory goodness. This assessment involves making judgments of the best of competing explanations. This abductive theory of method can serve as a useful framework for locating a number of more specific research methods within its fold.

Book Overview and Chapter Summary

This book undertakes a critical, in-depth examination of a selection of well-known, or otherwise important, quantitative research methods that are, or can be, used in behavioral science research. The book is interdisciplinary in nature and draws from varied literatures in research methodology, including the philosophy of science and statistical theory. As such, it is intended to serve as a useful complement to other books in the *Understanding Statistics* series. For example, the conceptual treatment of the method of exploratory factor analysis offered in Chapter 6 of the present book fits well with Fabrigar and Wegener's (2012) book in the series, *Exploratory Factor Analysis.*

In writing this book, my primary goal is to examine the conceptual foundations of a range of behavioral research methods. Some of them are well known. Others are seldom considered by behavioral science methodologists and researchers. A conceptual understanding of those methods is facilitated by presenting them in

relation to prominent accounts of scientific method where appropriate. The critical nature of the book is a natural consequence of dealing squarely with a conception of research methodology that is sponsored by the philosophy of scientific realism.

Chapter 1: This introductory chapter provides key ideas that should help make sense of the treatment of the five methods dealt with in the book. It begins by highlighting the importance and relevance of philosophy of science for understanding quantitative methods. It then gives a brief overview of the prominent philosophy of scientific realism, with particular emphasis on the nature of scientific methodology. After that, three major theories of scientific method are sketched because they figure in some of the ensuing chapters. Finally, an overview of the book's contents is provided before providing a note to the reader.

Chapter 2 focuses mainly on the nature, role, and importance of exploratory data analysis in behavioral research, although some attention is also given to the companion movement of computer-intensive statistics and its use of a reliabilist approach to justifying the knowledge claims it produces. Four perspectives on exploratory data analysis are presented, as they are shaped by different accounts of scientific method. One of these, the abductive theory of scientific method, locates exploratory data analysis in a multistage model of data analysis. Finally, John Tukey's outline of a philosophy for teaching data analysis is presented as an important part of an overall philosophy of exploratory data analysis.

Regarding Chapter 3, although widely used in behavioral science research, tests of statistical significance are poorly understood. In this critical examination of tests of significance, I discuss the questionable use of a popular hybridized form of significance testing in psychological research before outlining two plausible views of tests of significance: the neo-Fisherian and error statistical perspectives. These are judged to be superior to the hybrid version, especially that sponsored by the error-statistical account, which is a coherent philosophy of statistics. It is suggested that tests of significance can play a useful, if limited, role in research.

The subject of Chapter 4 is Bayesianism, which comprises both a philosophical theory of scientific confirmation and an influential perspective on statistics. I describe the nature of Bayesian

confirmation theory and assess its difficulties. I then compare it to two rivals: the hypothetico-deductive method and inference to the best explanation. In addition, I present, and evaluate, a neo-Popperian philosophy of Bayesian statistics, which is offered as an alternative to standard Bayesian modeling practice.

The primary concern of Chapter 5 is with the conceptual foundations of meta-analysis. The examination centers on large-scale issues having to do with meta-analysis and the nature of science. I give considerable space to presenting the conception of inquiry embodied in the underlying rationale of Gene Glass's approach to meta-analysis. I then examine David Sohn's provocative argument that meta-analysis is not a proper vehicle of scientific discovery. After that, I consider the role of meta-analysis in relation to the different processes of phenomena detection and scientific explanation. In doing so, I examine the extent to which meta-analysis can properly be said to contribute to scientific progress.

Chapter 6 examines the logic and purpose of exploratory factor analysis. It is argued that the common factors of exploratory factor analysis are not fictions, but latent variables best understood as genuine theoretical entities. This realist interpretation of factors is supported by showing that exploratory factor analysis is an abductive generator of elementary theories that exploits an important heuristic of scientific methodology known as the *principle of the common cause*. The importance of exploratory factor analysis is affirmed, and it is argued that it can be usefully combined with confirmatory factor analysis.

The concluding Chapter 7 assembles a number of important lessons learned from the preceding chapters before emphasizing the need for further work in the philosophy of research methods.

A Note for the Reader

The books in the *Understanding Statistics* series are fairly short in length. Thus, because I wanted to examine each method in this book in some conceptual detail, a limited number of methods were selected for consideration. Briefly, the reasons for my selection were as follows: Exploratory data analysis has a major role in pattern detection, and despite being advocated by the eminent statistician John Tukey for more than 50 years, it has not found a regular

place in the behavioral research method curriculum. The same can be said for computer-intensive resampling methods, which only arrived on the scene with the advent of high computing power. The reasons for the selection of tests of statistical significance should be obvious. They are overused in behavioral research and yet are poorly understood. It is high time that these failings were put right. The Bayesian approach to quantitative thinking earns its place in the book because it stands as the best-known theory of scientific confirmation, as well as the major rival school of thought to frequentist tests of statistical significance. Although the advocacy of Bayesian methods is on the rise, they also fail to figure regularly in the methods curriculum. Meta-analysis is a comparatively recent development in scientific methodology, but it has quickly become the dominant approach to reviewing primary empirical studies in the behavioral sciences. However, its conceptual foundations are only occasionally addressed. Finally, the long-standing method of exploratory factor analysis has been widely used in psychology, and many other sciences, for more than 70 years. Yet, its deep structure is seldom considered by those who describe it. It stands as our best example of a method that is well suited to the generation of explanatory hypotheses and theories.

The amount of space given to a description of each method varies considerably. Well-known methods, such as tests of statistical significance and exploratory factor analysis, receive little exposition. Less well-known methods, such as exploratory data analysis and resampling methods, receive more. Partly for this reason, and partly because of the paucity of a philosophical literature on these methods, their treatment in this book will seem less philosophical.

Finally, none of the methods considered in the book receives a full examination of its conceptual foundations. Relatedly, the reader should not expect from this book a series of definitive assessments about how one should understand and use the different methods. The book was written primarily as a stimulus for the reader to develop personal thinking about the methods considered in a manner that goes beyond what is contained in usual book presentations. The "Further Reading" section provided for each chapter should help the reader to extend thinking well beyond what the chapters themselves contain.

References

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis.* New York, NY: Oxford University Press.

Grice, J. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. New York, NY: Academic Press.

Greenwood, J. D. (1992). Realism, empiricism, and social constructionism. *Theory and Psychology*, *2*, 131–151.

Haig, B. D. (1987). Scientific problems and the conduct of research. *Educational Philosophy and Theory*, *19*, 22–32.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge, MA: MIT Press.

Hooker, C. A. (1987). *A realistic theory of science*. New York, NY: State University of New York Press.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, IL: University of Chicago Press (originally published 1962).

Manicas, P. T., & Secord, P. F. (1983). Implications for psychology of the new philosophy of science. *American Psychologist*, *38*, 399–413.

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, *48*, 1–30.

Mulaik, S. A. (1985). Exploratory statistics and empiricism. *Philosophy of Science*, *52*, 410–430.

Nickles, T. (1981). What is a problem that we might solve it? *Synthese*, *47*, 85–118.

Nickles, T. (1987). 'Twixt method and madness. In N. J. Nersessian (Ed.), *The process of science* (pp. 41–67). Dordrecht, the Netherlands: Nijhoff.

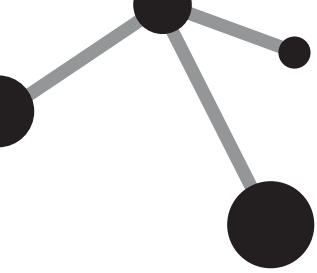
Proctor, R. W., & Capaldi, E. J. (2001). Empirical evaluation and justification of methodologies in psychological science. *Psychological Bulletin*, *127*, 759–772.

Psillos, S. (1999). *Scientific realism: How science tracks the truth*. London, England: Routledge.

Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ: Princeton University Press.

Trout, J. D. (1998). *Measuring the intentional world: Realism, naturalism, and quantitative methods in the behavioral sciences*. New York, NY: Oxford University Press.

Yu, C. H. (2006). *Philosophical foundations of quantitative research methodology*. Lanham, MD: University Press of America.



2

EXPLORATORY DATA ANALYSIS

> Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught.
>
> —J. W. Tukey, 1980

> [T]he time is ripe for a broad conceptualization of data analysis that includes the principles and procedures of EDA.
>
> —J. T. Behrens, 1997

Introduction

During the last 80 years, data analysis in statistics has placed its major emphasis on classical statistical inference, where the primary goal is to find out whether a set of data exhibits a designated feature of interest, associated with a probabilistic model. Such an approach to data analysis favors confirmatory research in which hypotheses are tested using methods such as tests of statistical significance (the topic of Chapter 3). Unfortunately, the dominance of this data analytic practice has had the effect of discouraging the genuine exploratory examination of data sets in terms of their quality and structure. Detailed explorations of data are essential

for detecting patterns, and it makes good sense to undertake such explorations instead of, or before, a probabilistic model is formulated and adopted.

However, since the early 1960s, a new empirical approach to data analysis in statistics has emerged. One important part of this development is *exploratory data analysis* (EDA), a process in which data are examined to reveal potential patterns of interest (e.g., Tukey, 1977, 1980). Another important part of this empirical data analytic movement is the advent of *computer-intensive resampling methods*, which, through repeated sampling of observed data, are used to produce a reference distribution (Efron & Tibshirani, 1993). These modern confirmatory resampling methods, rather than traditional confirmatory methods, can be employed as suitable complements to exploratory methods.

In 1968, the prominent statistician and scientist John Tukey gave an invited talk to the annual convention of the American Psychological Association; the talk was published in the association's flagship journal, the *American Psychologist*, the following year (Tukey, 1969). In his article, "Analyzing Data: Sanctification or Detective Work?" Tukey argued for the need to practice EDA, in the manner of a detective, as well as the confirmatory, or judicial, mode.

Unfortunately, psychology has, for the most part, ignored Tukey's advice, preferring instead to continue its focus on classical confirmatory methods. A comprehensive survey of all psychology PhD programs in the United States and Canada in 1986 revealed that only 20 percent of introductory statistics courses gave in-depth coverage to the topic of EDA (Aiken, West, Sechrest, & Reno, 1990). A replication and extension of that survey in the late 1990s contained no explicit information on this topic, though it did include coverage of "modern graphical displays" at 10 percent. An informal inspection of current standard textbooks on statistical methods in psychology shows that they give little attention to the topic of EDA. Occasional prominent calls in psychology recommending greater use of exploratory data analytic methods (e.g., Behrens, 1997; Wilkinson & the Task Force on Statistical Inference, 1999) seem to have made little difference. Clearly, the use of traditional confirmatory methods in data analysis remains the dominant practice.

This chapter is concerned with modern data analysis. It focuses primarily on the nature, role, and importance of EDA, although it gives some attention to the companion topic of computer-intensive

confirmatory methods. Because exploratory data analytic and computer-intensive resampling methods are less well known to behavioral scientists than the methods considered in other chapters, they receive more expository attention here than do key methods from other chapters. Considerable attention is also given to somewhat different perspectives on data analysis as they are shaped by four different accounts of scientific method. Before concluding, the chapter offers a brief presentation and discussion of Tukey's valuable, but underappreciated, philosophy of teaching data analysis.

It should be pointed out that the present chapter focuses on EDA in the manner of Tukey, given that such a view of EDA has not yet found its way into the modern behavioral science methods curriculum. The chapter does not consider the more recent exploratory data analytic developments, such as the practice of statistical modeling, the employment of data-mining techniques, and more flexible resampling methods (see, e.g., Yu, 2010).

What Is Exploratory Data Analysis?

In his landmark article, "The Future of Data Analysis" (Tukey, 1962), Tukey introduced the term *data analysis* to distinguish applied statistical work from the then-dominant formal inferential statistics. He characterized data analysis broadly in the following words:

> Data analysis . . . I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise, more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.
>
> Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole. Large parts of data analysis are incisive, laying bare indications which we could not perceive by simple and direct examination of the raw data, but these too are parts, not the whole. Some parts of data analysis . . . are allocation, in the sense that they guide us in the distribution of effort. . . . data analysis is a larger and more varied field than inference, or incisive procedures, or allocation. (Tukey, 1962, p. 2)

For Tukey, data analysis is both an empirical science and an art; it is not mathematical statistics (Tukey, in fact, advised serious students of data analysis that it was better that they aspire to be first-rate scientists rather than second-rate mathematicians); data analysis places a heavy emphasis on judgment, and it accepts satisfactory, not optimal, solutions to problems. Adopting an appropriate set of attitudes is an important part of Tukey's approach to data analysis. These attitudes include investigating realistic problems, making liberal use of ad hoc informal procedures in exploratory work, accepting the approximate nature of results, employing procedures iteratively, and including both exploratory and confirmatory approaches in the same analyses.

As the intellectual progenitor of modern EDA, Tukey developed a distinctive perspective on the subject that has helped to highlight its importance to research. It deserves to be considered as a philosophy of EDA in its own right and, more broadly, as a philosophy of data analysis (Dempster, 2003). Therefore, this brief examination of the philosophy of EDA pays particular attention to Tukey's thinking on the topic. As will be seen, Tukey took EDA to include more than descriptive statistics. Although both EDA and descriptive statistics are concerned with the visual display of data, EDA is also concerned with recognizing patterns in the data, tentatively forming hypotheses, and adopting the attitude of a detective when analyzing data.

According to Tukey (1980), data analysis should be treated as a two-stage, compound process in which the patterns in the data are first suggested by EDA and then critically checked through the use of confirmatory data analytic procedures. As already noted, EDA involves descriptive, and frequently quantitative, detective work designed to reveal structure or pattern in the data sets under scrutiny. The data analyst is encouraged to undertake an open-eyed investigation of the data and perform multiple analyses using a variety of intuitively appealing and easily used techniques.

The compendium of methods for the exploration of data, many of which were developed by Tukey (1977), sometimes in association with colleagues, is designed to facilitate both discovery and communication of information. These methods are concerned with the effective organization of data, the construction of graphical and semigraphical displays, and the examination of distributional assumptions and functional dependencies. As noted further in this chapter, two additional attractive features of Tukey's methods are their resistance to changes in underlying

distributions and their resistance to outliers in sets of data. Exploratory methods with these two features are particularly suited to data analysis in psychology and the behavioral sciences, where researchers are frequently confronted with ad hoc sets of data on amenable variables, which have been acquired in convenient circumstances.

In the next section, I briefly describe the two best known methods of EDA—stem-and-leaf displays and box-and-whisker plots—and then provide an overview of the four focal themes of EDA: resistance, residuals, re-expression, and revelation.

Two Methods of Exploratory Data Analysis

The Stem-and-Leaf Display

The stem-and-leaf display is a clever, useful, and easily understood device that can provide helpful first impressions of a set of data. The stem-and-leaf display economically organizes values from a set of data in numerical order, while visually displaying the shape, spread, and distributional characteristics in the manner of a histogram. Unlike a histogram, however, the stem-and-leaf-display can be readily constructed by hand, and unlike most displays, it has the virtue of retaining information on individual data values. A stem-and-leaf display will decompose each value into a stem value and a leaf value. For example, for an ordered set of data values, each tens digit would comprise the vertical stem column, while the units digit would be arrayed horizontally as the leaf values. Stem-and-leaf displays can also be placed back to back to compare similar batches of data for detailed differences.

The Box-and-Whisker Plot

The box-and-whisker plot, sometimes simply called the *box plot*, is also helpful in the presentation of data. The box plot is a graph comprising rectangular boxes and lines that display the distributional shape, central tendency, and variability of a set of observations. It is particularly helpful for indicating skewness and the presence of outliers in a distribution. From a box plot we can obtain the "five-number summary" of a set of data. This summary comprises the minimum and maximum values within a set, along with the median and the lower and upper quartiles. In effect, the

box plot provides an effective graphical display of the five-number summary. Because of this, the box plot is particularly useful for comparing two or more sets of data.

Of course, there are many more methods suitable for EDA, but the two just noted should give a sense of how semigraphical methods, which combine both graphical and tabular information, help meet the primary goals of exploring data.

The Four Rs of Exploratory Data Analysis

Tukey and his collaborators (Hoaglin, Mosteller, & Tukey, 1983) argued that the tools of EDA can be organized according to four central themes: resistance, residuals, re-expression, and revelation. These themes usefully help to distinguish EDA from classical inferential statistics.

1. ***Resistance.*** An analysis, or summary, of the data is resistant if it is not sensitive to misbehaving or unusual data. This requirement of EDA is not met when unexpected data values mislead the summary of the bulk of the data, as happens when data misbehave, are unusual, or are otherwise misleading. These situations occur often enough to make resistance an important consideration in EDA. For example, in EDA, the median is the most commonly used measure of central tendency because it is not affected by outliers as much as the mean; it is a more resistant, or robust, statistic.
2. ***Residuals.*** The careful examination of residuals is important in EDA. They are the deviations of observations from the value predicted by a tentative model. In EDA (and in confirmatory data analysis), residuals are summarized to obtain a sense of overall fit. Stated differently, EDA uses the framework

   $$\text{data} = \text{fit} + \text{residual}$$

   Hence, data residuals are what remain of the data after a summary or the fitted model has been subtracted, according to the equation

   $$\text{residual} = \text{data} - \text{fit}$$

As leftover parts of the data, residuals still need attention, which may involve a reanalysis and refit.

3. ***Re-expression.*** *Re-expression* is the term used in EDA for rescaling or transforming data. Re-expression is common in EDA because data are often collected in a manner based on convenience, or habit, rather than careful attention to scaling. EDA looks to find more easily interpreted scales through re-expression. Re-expressions that lead to symmetric distributions are preferred for they promote the interpretation of general linear models, improve comparisons across groups, and reflect the structure of the sampling distributions. In psychology, logarithmic, arcsine, and reciprocal transformations are often used.
4. ***Revelation (Display).*** Displays are often graphical (and semigraphical) in nature. They allow the data analyst to see the behavior of the data and, as the analysis proceeds, also the behavior of the residuals and various diagnostic measures. EDA emphasizes the frequent use of displays to ensure that unexpected features of the data are not overlooked. Traditional EDA emphasizes a number of relatively simple numerical and graphical techniques for displaying data. For example, box-and-whisker plots are often used to represent univariate data. However, significant advances in graphics and data visualization in more recent times have enabled data analysts to construct all manner of revealing displays of complex phenomena (e.g., Chen, Härdle, & Unwin, 2008).

Psychology's attitudes toward each of the four Rs deserve comment. Checking for resistance, or robustness, of methods to violation of their assumptions with sets of data is done much less frequently than should be the case. The framework, residual = data – fit, is widely employed in psychology, but almost entirely in the context of confirmatory data analysis (I comment further on Tukey's reluctance to accord models a role in EDA). Re-expression of data through transformation is frequently done in psychology, although typically with limited knowledge of prevalence of different types of distribution in its varied subject domain

(Micceri, 1989). Finally, the importance of revelation, or data display, in EDA affirms the adage that a picture is worth a thousand words, an insight that is founded on the fact that humans (and other vertebrates) are primarily visual creatures (Gould, 1994). Recently, psychology has given more attention to the value of good data displays (e.g., Lane & Sándor, 2009), although the major focus is seldom on EDA.

I turn now to an examination of the philosophical foundations of EDA by considering the role it plays in four different accounts of scientific method.

Exploratory Data Analysis and Scientific Method

In his writings on data analysis, Tukey emphasizes the related ideas that psychology is without an agreed-upon model of data analysis, and that we need to think more broadly about scientific inquiry. In his address to the American Psychological Association mentioned previously, Tukey (1969) presented the following anonymous excerpt from a prominent psychologist for his audience to ponder. I quote in part:

> I have the feeling that psychology is currently without a dominant viewpoint concerning a model for data analysis. In the forties and early fifties, a hypothetico-deductive framework was popular, and our mentors were keen on urging the design of "crucial" experiments for the refutation of specific predictions made from one or another theory. Inductive empiricism was said to be disorderly and inefficient. You and I knew then, as we know now, that no one approach is uniformly most powerful. (p. 90)

It is not surprising that mention is made of hypothetico-deductive and inductive inquiry in the quotation for these two outlooks are generally acknowledged as the two most influential conceptions of scientific method in the history of science (Laudan, 1981). In what follows, I consider EDA in relation to the hypothetico-deductive and inductive methods of science. I then discuss Tukey's proposed framework of inquiry, which he believes properly accommodates EDA, before considering EDA in relation to an abductive theory of scientific method. My treatment should be understood as endorsing the assertion in the quotation that there is no one account of scientific method that is best for all occasions.

Exploratory Data Analysis and the Hypothetico-Deductive Method

According to the standard conception of the hypothetico-deductive method, a scientist takes a hypothesis or a theory and tests it by deriving from it one or more observational predictions, which are amenable to a direct empirical test. If the predictions are borne out by the data, then that result is taken as a confirming instance of the theory. If the predictions fail to square with the data, then that fact counts as a disconfirming instance of the theory.

Most psychological researchers continue to undertake their research within the confines of this conception of hypothetico-deductive method. Witness their heavy preoccupation with theory testing, where confirmatory data analyses are conducted on limited sets of data gathered in accord with the dictates of the test predictions of theories. In this regard, psychologists frequently employ tests of statistical significance to obtain binary decisions about the credibility of the null hypothesis and its statistical or substantive alternative (see Chapter 3). However, the heavy use of tests of statistical significance in this way strongly discourages researchers from looking for more interesting patterns in the data of potential interest. Indeed, the continued neglect of EDA in psychological research occurs in good part because there is no acknowledged place for such work in the hypothetico-deductive conception of inquiry (Wilkinson & the Task Force on Statistical Inference, 1999). It is important to understand that the hypothetico-deductive method itself is not at fault here for it is a confirmatory procedure, not an exploratory procedure. Rather, it is the dominating use of the hypothetico-deductive method by researchers that clouds their ability to appreciate EDA as an important element of scientific research.

Exploratory Data Analysis and the Inductive Method

The most popular characterization of inductive scientific method maintains that inquiry begins by securing observed facts, which are collected in a theory-free manner. These facts provide a firm base from which the scientist reasons "upward" to hypotheses, laws, or theories. The reasoning involved takes the form of enumerative induction and proceeds in accordance with some

governing principle of inductive reasoning. This rather simple view of inductive method can be defended in a moderate form. In psychology, the radical behaviorism of B. F. Skinner (1984) makes use of a nonstatistical inductive conception of scientific method. The major goals of Skinner's conception of inductive method are, first, to detect empirical generalizations about the subject matter of interest and then to systematize those empirical generalizations by assembling them into nonexplanatory theories.

I think the worth of the inductive method as a model for data analysis is dismissed too quickly in the previous quotation from Tukey (1969). The major limitation of the inductive account of scientific method lies not so much with its perspective on data analysis but with the prohibition of the formulation of explanatory theories by many of its proponents. It will be seen shortly that a conception of inductive method is embedded in the broader abductive account of scientific method.

Exploratory Data Analysis and Tukey's Model of Inquiry

Tukey (1980) believes that statisticians and data analysts have given too little attention to broad concerns about inquiry. He maintains that much data analysis proceeds according to the following linear conception of confirmatory research:

$$\text{Question} \rightarrow \text{Design} \rightarrow \text{Collection} \rightarrow \text{Analysis} \rightarrow \text{Answer}$$

Tukey argues that this model of inquiry is incomplete and that it neglects the following set of questions and answers, all of which have to do with data exploration. He writes:

1. How are questions generated? (Mainly by quasi-theoretical insights and the exploration of past data.)
2. How are designs guided? (Usually by the best qualitative and semiquantitative information available, obtained by exploration of past data.)
3. How is data collection monitored? (By exploring the data, often as they come in, for unexpected behavior.)
4. How is analysis overseen; how do we avoid analysis that the data before us indicate should be avoided? (By exploring the data—before, during, and after analysis—for hints, ideas, and sometimes, a few conclusions-at-5 *percent/k*.). (p. 23)

Tukey argues that, to pose and answer these questions, and indeed to implement properly confirmatory research, we need to reorganize the model of inquiry along the following lines:

Ides → Question →← Design → Collection → Analysis → Answer

This is to say, we "begin" not with a properly formulated question, but an idea of a question that cannot be given an answer until it is specified in terms of appropriate constraints. And, to do this requires exploration, which if we are successful, will lead to a circumscribed question that warrants attempted confirmation. For this reason, Tukey (1980) maintains that "finding the question is often more important than finding the answer" (p. 23).

Clearly, if science is to derive maximum benefit from data analysis, it needs to take seriously both its exploratory and its confirmatory modes. It has already been noted that Tukey regards data analysis as a two-stage compound process: exploratory followed by confirmatory. However, in his more precise moments, Tukey speaks of data analysis as a three-stage process, where the stages lie on a continuum of data analysis (Tukey, 1972): The first stage is that of EDA, where the investigator seeks to learn what is going on in the data. The second stage Tukey calls *rough confirmatory data analysis*. Here hypotheses are refined and rough tests are carried out, often using estimation techniques such as confidence intervals. In the third stage, known as *strict confirmatory data analysis*, the investigator tests well-specified hypotheses using modern robust statistical methods. Thus, it should be clear that, for Tukey, confirmatory data analysis is just as important as EDA. The heavy focus in his writings on EDA is more a function of their comparative neglect than their greater importance.

Exploratory Data Analysis and the Abductive Method

As stated earlier, in the abductive theory of method (Haig, 2005, 2014), scientific inquiry proceeds as follows: Guided by evolving research problems, sets of data are analyzed to detect robust empirical regularities, or phenomena. Once detected, these phenomena are explained by abductively inferring the existence of underlying causes that are thought to give rise to the phenomena. Here, abductive inference involves reasoning from claims about phenomena,

understood as presumed effects, to their theoretical explanation in terms of underlying causes. Upon positive judgments of the initial plausibility of these explanatory theories, attempts are made to elaborate on the nature of the causal mechanisms in question. This is done by constructing plausible models of those mechanisms by analogy to relevant ideas in domains that are well understood. When the theories are well developed, they are assessed against their rivals with respect to their explanatory goodness. This assessment involves employing criteria specifically to do with explanatory worth.

It should be apparent, even from this brief sketch, that the abductive theory is considerably broader than both the hypothetico-deductive and inductive alternatives (and all three accounts go beyond Tukey's near-exclusive concern with data analysis). The breadth of the abductive theory enables it to operate as a framework theory within which an extensive array of data analytic and theory construction methods and strategies can be usefully located.

Exploratory Data Analysis in a Multistage Model of Data Analysis

The important place of EDA in the abductive theory of method can be appreciated by describing its role in the process of phenomena detection. Phenomena are relatively stable recurrent general features of the world that we seek to explain (Haig, 2014; Woodward, 1989), and their detection frequently involves an inductive process of empirical generalization. This inductive process of phenomena detection reserves an important place for the exploratory analysis of data. In detecting phenomena, one is concerned to extract a signal from the noise of data, and for this, the intensive search of large amounts of data is frequently required. This is precisely because securing a heavy information yield from our data is likely to provide potentially interesting data patterns that might turn out to be genuine phenomena. In this context, data mining is encouraged, and the capabilities of exploratory techniques in this regard often make them the appropriate methods of choice.

A more fine-grained appreciation of the role of EDA in the discovery of empirical phenomena can be gained by outlining a multistage model of data analysis with EDA as one of its stages. This model, which is featured in the abductive theory of method, describes one of a number of ways in which empirical phenomena can be detected.

The model comprises the four sequenced stages of initial data analysis, EDA, close replication, and constructive replication.

Initial Data Analysis. The initial examination of data (Chatfield, 1985) refers to the first informal scrutiny and description of data that is undertaken before EDA proper begins. It involves screening the data for their quality. Initial data analysis variously involves checking for the accuracy of data entries, identifying and dealing with missing and outlying data, and examining the data for their fit to the assumptions of the data analytic methods used. Data screening thus enables one to assess the suitability of the data for the type of analyses intended. The initial analysis of data has much in common with Tukey's approach to EDA. However, these two related data analytic endeavors serve different primary functions (data screening and pattern detection, respectively), and I restrict initial data analysis to the preliminary scrutiny of data before exploratory EDA (in Tukey's sense) begins.

Exploratory Data Analysis. Given that EDA is the major focus of this chapter, it suffices to say here that it plays an indispensable role in the detection of patterns in data that are the springboard to the eventual discovery of phenomena, or robust empirical regularities.

Close Replication. Successfully conducted exploratory analyses will suggest potentially interesting data patterns. However, it will normally be necessary to check on the stability of the emergent data patterns though use of appropriate confirmatory data analysis procedures. Computer-intensive resampling methods such as the bootstrap, the jackknife, and cross validation (Efron & Tibshirani, 1993) constitute an important set of confirmatory procedures that are well suited to this role. They are briefly discussed in material that follows.

Constructive Replication. In establishing the existence of phenomena, it is often necessary to undertake both close and constructive replications. The statistical resampling methods just mentioned are concerned with the consistency of sample results that help researchers achieve close replications. By contrast, constructive replications are undertaken to demonstrate the extent to which results hold across different methods, treatments, and occasions. In other words, constructive replication is a triangulation strategy designed to ascertain the generalizability of the results identified by successful close replication (Lindsay & Ehrenberg,

1993). Constructive replication, in which researchers vary the salient study conditions, is a time-honored strategy for justifying claims about phenomena.

The four-stage model of data analysis just outlined assists in the detection of phenomena by attending in turn to the different, but related, tasks of data quality, pattern suggestion, pattern confirmation, and generalization. To repeat, the role of EDA in this process is that of pattern detection.

Exploratory Data Analysis and Abductive Inference

Having commented on the role of EDA in the process of phenomena detection from the vantage point of the abductive theory of scientific method, an important question remains: What is the nature of the relations between EDA and abduction? In psychology, I (Haig, 2013) and Behrens, Dicerbo, Yel, and Levy (2013) (see also Behrens & Yu, 2003) have commented on philosophical aspects of EDA. We hold contrasting views about the relevance of abductive reasoning as a core component of the philosophy of EDA. Behrens and his coauthors think abduction provides the "core logic" of EDA. I disagree. In this section, I say why I think their position is mistaken, and that their charge that mine is "a particularly disturbing" view of EDA (Behrens et al., 2013, p. 39) is unfounded.

Abduction as a form of inference is not well known in academic circles. Broadly speaking, abduction is concerned with the generation and evaluation of *explanatory* hypotheses. In this sense, it contrasts with the more familiar ideas of inductive and deductive inference. Behrens et al. (2013) begin by taking their cue from the philosopher-scientist Charles Peirce and state that abduction is the form of inference involved in generating new ideas or hypotheses. However, surprisingly, Behrens et al. do not stay with Peirce on this matter. Instead, they closely follow Josephson and Josephson (1994) and characterize abductive inference according to the following pattern of reasoning (p. 39):

> D is a collection of data (facts, observations, givens).
>
> Hypothesis H explains D (would if true, explain D).
>
> No other hypothesis explains D as well as H does.
>
> Therefore, H is probably correct.

Patently, this argument schema does not describe the abductive process of *hypothesis generation*. Instead, it characterizes the abductive form of reasoning known as *inference to the best explanation*. Inference to the best explanation is used in science to appraise competing theories in terms of their explanatory goodness (Thagard, 1992). For the schema to capture abductive hypothesis generation, the third premise, which refers to competing hypotheses, would have to be deleted, and the conclusion would be amended to say that the hypothesis in question was initially plausible, not probably correct.

Despite the fact that philosophers of science sometimes speak of abduction and inference to the best explanation as though they were the same thing, it is important to differentiate between the abductive generation of hypotheses and their comparative appraisal in terms of inference to the best explanation. They are discernibly different phases of theory construction. In short, Behrens et al. (2013) adopt a conception of abduction that is ill-suited to explicating the process of idea generation, whether it be pattern identification through EDA or some other generative process. As a result, they fail to make an instructive connection between their chosen characterization of abduction and the reasoning involved in EDA.

However, my major worry is not that Behrens et al. (2013) choose the wrong form of abduction to explicate the inferential nature of EDA, but that they try to understand it by appealing to abduction at all. The fundamental difference between our opposed views can be brought out by drawing, and adhering to, the important three-fold methodological distinction between data, phenomena, and explanatory theory. Briefly, data are idiosyncratic to particular investigative contexts, and they provide the evidence for phenomena, which are recurrent general features of the world that we seek to explain. In turn, phenomena are the appropriate source of evidence for the explanatory theories that we construct to understand empirical phenomena. I have just described one way of detecting phenomena by outlining a multistage model of data analysis. These stages of data analysis are concerned in turn with assessing data quality, detecting data patterns, confirming those patterns through use of computer resampling methods (a prominent feature of Tukey's conception of data analysis), and establishing the reach of the confirmed relationships in the form of inductive generalizations. Viewed in this context, EDA is an

empirical, descriptive, pattern detection process. It is one component in a sequence of activities that, if undertaken successfully, can lead to the detection of new empirical phenomena.

Once claims about empirical phenomena are established, there is a natural press to understand them by constructing one or more explanatory theories. It is here, and not with the process of phenomena detection, that abduction does its work. In other work (Haig, 2005, 2014), I argue how by different abductive means, one can generate explanatory theories, develop them through analogical modeling, and evaluate them in relation to their rivals in terms of inference to the best explanation. Importantly, the means I choose for showing this are, in turn, the abductive methods of exploratory factor analysis, analogical abduction, and the theory of explanatory coherence (Thagard, 1992). As methods, they provide rich abductive resources that enable researchers to produce explanatory knowledge. They well exceed the rudimentary account of abduction provided by the previous argument schema for inference to the best explanation.

Behrens et al. (2013) speak of generating hypotheses in the context of EDA. In this regard, they pose questions about things such as skewness and partialling out. Of course, these sorts of questions can be framed as hypotheses, but they are *descriptive hypotheses, not explanatory hypotheses*. They are hypotheses about data analytic matters; they are not explanations of the data patterns that result from exploratory data analytic work.

The Collected Works of John W. Tukey (Vols. 3 and 4; Jones, 1986) provide valuable information about Tukey's wide-ranging philosophy of data analysis, including EDA. I advocate an essentially Tukeyan philosophy of data analysis (Haig, 2013). This may surprise Behrens et al. (2013), who see my philosophy as opposed to Tukey's. However, I see no tension, let alone a contradiction, in subscribing to large parts of Tukey's perspective on data analysis on the one hand and advocating a thoroughgoing abductive perspective on theory construction on the other. This is made possible by taking the compendium of exploratory data analytic methods as true to their name (they are *data analytic* methods) and abductive methods as true to their name (they are methods concerned with the construction of *explanatory* hypotheses and theories).

If researchers were to follow Behrens et al. (2013) and characterize EDA as fundamentally abductive in nature, they would

risk construing descriptive hypotheses as explanatory hypotheses, when they had done no explanatory work at all. For this reason, I think it would be better to put abduction to one side and follow Tukey's philosophy of EDA.

Exploratory Data Analysis After Tukey

We have seen that in ushering in the empirical approach to data analysis, Tukey argued for the importance of EDA, developed many of its tools, and formulated a systematic perspective on the subject. However, Tukey's pioneering 1962 article underestimated the impact of the computer in modern data analysis and did not sufficiently acknowledge the relevance of modeling to the endeavor (Huber, 2011). Both of these matters deserve some comment.

Among other things, EDA for Tukey was a pencil-and-paper activity in which one uses graph paper and transparencies. Tukey himself had no need for the computer as an aid to calculation because he was highly adept at computation by just using his brain. However, everyone today necessarily practices EDA as a computer-assisted endeavor, both for constructing graphical displays and for computation. This is especially so because very large data sets are now being subjected to exploratory investigation.

Classical inferential statistics makes essential use of mathematical models of the data. By contrast, EDA, as understood by Tukey, does not. Tukey warns against the dangers of using models in the exploratory phase of data analysis. He maintains that the early use of models forces data into a procrustean bed and lessens the chances of detecting potentially interesting patterns. In addition, he thinks that premature modeling makes it more difficult to detect, and make sound judgments about, when outliers in the data can be ignored. For Tukey, the use of mathematical models is guided by EDA and comes into play in the subsequent confirmatory phase. In short, Tukey is not opposed to the use of mathematical models, only their use in the early exploratory phase. Interestingly, Lenhard (2006) suggests that EDA does in fact adopt a nonmathematical, and more instrumental, conception of models, but I shall not pursue that line of thought here.

Finally, I comment on the relation between EDA and Bayesian statistics, bearing in mind that it is commonly thought that EDA and Bayesian inference are unrelated statistical endeavors. For his part, Tukey expresses reservations about Bayesian data analysis. Although he thinks it would be a mistake to discard Bayesian analyses altogether (he acknowledges that in restricted contexts Bayesian analyses might be of some use in bringing in information not contained in the data), he is opposed to its widespread use as a highly formalized and unified, "seemingly scientific," process.

However, the Bayesian philosopher and statistician I. J. Good (1983) has sketched a philosophy of EDA that endorses many of the aims of EDA, as spelled out by Tukey. However, unlike Tukey, Good is willing to appeal to subjective prior probabilities to help judge whether patterns in the data are potentially explicable in the particular context within which they arise. Good recognizes the very sketchy nature of his philosophy of EDA and views his effort as groping toward what is needed. The question of whether one should make subjective estimates of prior probabilities is discussed in Chapter 4.

Finally, in this section, I note that Andrew Gelman (2004), whose distinctive philosophy of Bayesian inference is discussed in Chapter 4, endeavors to bring EDA and Bayesian inference together by showing how the former can be embedded into the probability-modeling outlook of the latter. Essentially, his view is that EDA and Bayesian inference complement each other. He says that "(a) exploratory and graphical methods can be especially effective when used in conjunction with models, and (b) model-based inference can be especially effective when checked graphically. Our key step is to formulate essentially all graphical displays as model checks, so that new models and new graphical methods go hand-in-hand" (Gelman, 2004, p. 757).

Resampling Methods

The Computer-Intensive Philosophy

The classical statistical methods developed by Fisher, Neyman and Pearson, and others of their time were tailored to the limited calculational abilities of the human mind and the mechanical calculator and had to be augmented by appropriate mathematical analysis.

However, since the 1980s, statisticians have been able to exploit the massive computational power of the modern computer and develop a number of computer-intensive resampling methods, such as the jackknife, the bootstrap, and cross validation (Efron & Tibshirani, 1993). These methods constitute one important set of confirmatory procedures that are well suited to the task of checking on the data patterns evoked by EDA. By exploiting the computer's computational power, these resampling methods free us from the restrictive assumptions of modern statistical theory, such as the belief that the data are normally distributed, and permit us to gauge the reliability of chosen statistics by making thousands, even millions, of calculations on many data points.

In the second half of this chapter, I briefly describe three prominent resampling methods and briefly address some issues that arise from consideration of their conceptual foundations. Some of these are helpfully discussed in Yu (2003, 2008) and Sprenger (2011). Resampling methods comprise a broad family, and there are a number of variants of each of these procedures.

The Jackknife

Perhaps the best known computer-intensive resampling method is the jackknife, a method that was introduced by Quenouille (1949) and developed early by Tukey and his associates (e.g., Mosteller, 1971; Tukey, 1958), who showed that it could improve the variance as well as the bias of an estimate. In everyday contexts, the jackknife is a scout's single-bladed knife that can be used tolerably well for a variety of purposes. By analogy, the term *jackknife* was adopted by Tukey to suggest a single confirmatory statistical tool that could be applied in a wide variety of situations when specialized tools were not available. In fact, Tukey recommends using the jackknife as an all-purpose method when seeking the confirmation of data patterns initially suggested by exploratory methods. The jackknife is an important attempt to establish the accuracy of a computed estimate of some quantity of interest, such as a mean, a standard deviation, or a correlation. It proceeds by removing one observation at a time from the original data set and recalculating the statistic of interest for each of the data sets. The variability of the statistic across all of the truncated data sets can then be described by giving us

an empirically obtained measure of the reliability or stability of the original estimate. The jackknife is not massively computationally intensive in that the computational work it requires can often be done with a calculator. Although the jackknife is still used, it is used less often than the bootstrap, although the two methods can be used in tandem in particular contexts.

The Bootstrap

A method closely related to the jackknife is the bootstrap. This more recent method holds considerable promise as a powerful and flexible, data-based method for statistical inference. Numerical answers are obtained by employing computer algorithms rather than tractable mathematical theory, though mathematical statistics now lies behind bootstrap thinking.

The bootstrap has been primarily developed by Bradley Efron and his colleagues (Diaconis & Efron, 1983; Efron, 1979; Efron & Tibshirani, 1993). Efron gave the method its name by likening it to the familiar idea of making progress by lifting oneself by the bootstraps. The bootstrap in statistics embodies a self-help algorithm that enables the researcher to create many samples from the data available in one given sample. Early development of the bootstrap was a straightforward extension of the jackknife, but the bootstrap now stands as a general-purpose method that can be used to tackle a wide variety of statistical estimation problems. Computationally speaking, the bootstrap sampling distribution is usually constructed via Monte Carlo simulation, in which a number of samples, say 100 or 1,000, are drawn from the observed data of the available sample rather than from some hypothetical distribution, as is the case with classical statistics. Repeated samples of the same size as the observed sample with replacement from the data are drawn, the chosen statistic of each bootstrap sample is computed, and the variance of this set of means is calculated.

Three major benefits of the bootstrap are that it involves no distributional assumptions, it has a wide variety of applications, and it can be used with small as well as large samples. Additionally, despite claims that the bootstrap is a nonparametric procedure, it can in fact be applied both parametrically and nonparametrically.

Cross Validation

The basic idea of cross validation has long been considered important to science and is familiar to many psychologists. It is often used to assess the adequacy of a statistical model. At a minimum, this time-honored strategy typically involves randomly splitting the data into two subgroups of approximately equal size. The results of interest are calculated from the data of one subgroup, after which their confirmation is sought by comparing them with the results of the other subgroup. This long-standing idea has been taken up in a computationally intensive way, with curves fitted to one half of the data and then validated by testing successively for the best fit to the second half. The data do not have to be split in half and at random; they can be split many times and in many different ways. Alternatively, cross validation can be used in the manner of the jackknife by leaving out a single observation at a time, or a data set can be split into a number of subsets, with each subset being left out in turn as the validation set.

Resampling and Counterfactual Reasoning

One criticism sometimes leveled at the bootstrap is that its name suggests the idea of a procedure with which researchers try to obtain something for nothing, which cannot be achieved—in this case, an empirical sample obtained by an appropriate sampling procedure. The corresponding image is of "statisticians idly resampling from their samples, presumably having about as much success as they would if they tried to pull themselves up by their bootstraps" (Hall, 1992, p. 2). However, this image is misleading because the bootstrap metaphor contains a nonlinear logic that is often used in science to good effect. Realistically, the boot pushes against the hand, while its straps offer the researcher lift-off assistance. In the case of statistical bootstrapping, the original sample pushes upward, while the resampling strategy helps lift the investigation to the obtained empirical sample. Moreover, simulation studies indicate that sound conclusions can often be reached through use of bootstrap procedures, but being entirely conditioned by the sample data, they do not guarantee success (Hall, 1992).

It is relevant to note here that scientists often make explicit use of "what if" counterfactual thinking to better understand the

actual world. Counterfactuals are contrary-to-fact conditional statements, which take the form, "If *p* hadn't happened, then *q* wouldn't have happened," or "If *p* has happened, then *q* would have happened." Philosophers also make frequent use of counterfactual thinking to analyze important metascientific concepts such as causation, laws, and dispositions.

As a statistician and scientist, Fisher made deliberate use of counterfactual thinking. For example, his appeal to theoretical distributions, a long-run frequentist conception of probability, and the construction, and use of, the randomization or permutation test all involve counterfactual reasoning in the sense that they contain elements that do not map onto the real world. It is relevant to note that Fisher specifically appreciated the worth of empirically generated sampling distributions, as seen with his development of the randomization test, which was a forerunner to the resampling methods described previously. In fact, he went to great lengths to construct an empirical sampling distribution, despite the rudimentary computational resources at his disposal. It seems likely that with sufficient computing resources, Fisher would have made considerable use of empirically generated sampling distributions, rather than relying on theoretical sampling distributions, such as the *F* and *t* distributions (Rodgers, 1999).

Reliabilist Justification

It is important to appreciate that the resampling methods just mentioned make use of an approach to justification known as *reliabilism* (Goldman, 1986). Here, the reliability checks on emergent data patterns are provided by the consistency of test outcomes, which is a time-honored validating strategy. Our willingness to accept the results of such checks is in accord with what Paul Thagard (1992) calls the *principle of data priority.* This principle asserts that statements about observational data, including empirical generalizations, have a degree of acceptability on their own without being bolstered by a theory that explains them. Such claims are not indubitable, but they do stand by themselves better than claims justified solely in terms of what they explain. What justifies the provisional acceptance of data statements is that they have been achieved by reliable methods. Specifically, what strengthens our provisional belief in the patterns elicited

by EDA is their confirmation through use of computer-based resampling methods.

Further, it is important to appreciate that the acceptability of claims provided by the reliabilist justification of computer-intensive resampling methods can be enhanced by making appropriate use of what is called a *coherentist* approach to justification. One important form of coherence is explanatory coherence, and one method that delivers judgments of explanatory coherence is the theory of explanatory coherence (Thagard, 1992). According to this theory, data claims, including empirical generalizations, receive an additional justification if and when they enter into, and cohere with, the explanatory relations of the theory that explains them.

Computer-intensive resampling methods have been slow to catch on in applied statistics, and their presence on the contemporary psychological landscape is characterized by little more than the occasional demonstration paper, with applications and tutorials (e.g., Lunneborg, 1985; Thompson, 1991; Yu, 2008). However, with the increasing availability of suitable software for the implementation of these methods, it is to be hoped that their introduction to statistics education will soon see them become part of the behavioral scientist's standard toolkit. Now that psychology seems finally poised to officially embrace EDA, we can hope for a corresponding increase in the use of modern confirmatory statistical methods that will enable us to ascertain the validity of the data patterns initially suggested by use of exploratory methods.

A Philosophy for Teaching Data Analysis

An underappreciated, but important, feature of Tukey's writings on EDA is the illuminating remarks on the teaching of data analysis that they contain. These remarks can be assembled into a sketch of an instructive philosophy for teaching data analysis, which can properly be regarded as part of an overall philosophy of EDA. Tukey's philosophy of teaching advises us to think about and teach data analysis in a way that is quite different from the prevailing custom.

Provocatively, Tukey (1980) maintains that the proper role of statistics teachers is to teach that which is most difficult and leave that which is more manageable to good textbooks and computers. He recommends teaching data analysis the way he understands biochemistry was taught—concentrating on what the discipline of

statistics has learned, perhaps with a discussion of how such things were learned. The detail of methods should be assigned to laboratory work, and the practice of learning data analytic techniques should be assigned to a different course in which problems arose. Tukey foresaw that such a redirection in teaching data analysis would have to be introduced in phases. In Tukey's (1962) words, "the proposal is really to go in the opposite direction from cookbookery; to teach not 'what to do,' nor 'how we learned what to do,' but rather, 'what we have learned' " (p. 63). This advice is broadly consistent with the idea raised in the book's introduction, that we should teach research methods in terms of their accompanying methodology.

Another prominent feature of Tukey's philosophy of teaching data analysis is his recommendations that we should teach both exploratory and confirmatory data analysis and that we have an obligation to do so. Tukey's strong promotion of the value of EDA was intended as a counter to the dominance of confirmatory data analysis in statistical practice. However, as already noted, for Tukey, EDA was not to be understood as more important than confirmatory data analysis because both are essential to good data analysis.

Tukey also suggests that EDA should probably be taught before confirmatory data analysis. There are several reasons why this recommendation makes good sense. Properly taught, EDA is probably easier to learn, and it promotes a healthy attitude to data analysis (encouraging one to be a dataphile without becoming a data junkie). It requires the investigator to get close to the data, analyze the data in various ways, and seek to extract as much potentially important information from them as possible. This is done to detect indicative patterns in the data before establishing through confirmatory data analysis that they are genuine patterns.

Tukey emphasizes that learning EDA centrally involves acquiring an appropriate attitude toward the data, which includes the following elements: EDA is sufficiently important to be given a great deal of time; EDA should be carried out flexibly with multiple analyses being performed (there is no one best analysis of the data); and EDA should employ a multiplicity of methods that enhance visual display.

Conclusion

Although data analysis has become an important part of professional statistics, both exploratory and computer-intensive resampling methods remain a minority practice in psychology. Given the importance of both to the field of data analysis, psychological research would benefit by placing as much emphasis on them as they do classical confirmatory methods. Tukey's insistence that EDA should precede confirmatory data analysis, and that confirmatory data analysis should feature resampling methods, is sound advice. This chapter has given considerable attention to the place of data analysis in different conceptions of scientific method, in particular the abductive theory of method. Specifically, the location of EDA in its four-stage model of data analysis shows one way in which it can contribute to the important scientific process of phenomena detection. Finally, Tukey's unheralded philosophy of data analysis, including his philosophy of EDA, offers the best articulation of the conceptual foundations of data analysis that has been expressed by one voice. The field of data analysis would be conceptually enriched by heeding Tukey's philosophical contributions to the topic and working through their implications for thinking about, and practicing, data analysis.

Further Reading

John Tukey's groundbreaking book, *Exploratory Data Analysis* (Boston, MA: Addison-Wesley, 1977), is the major text on EDA. It stimulated much further work in the field of data analysis.

Volumes 3 and 4 of *The Collected Works of John W. Tukey* (Pacific Grove, CA: Wadsworth & Brooks/Cole, 1986) bear the name *Philosophy and Principles of Data Analysis*. The 30 articles contained in the two volumes contain valuable information about Tukey's wide-ranging philosophy of data analysis.

A. P. Dempster ("John W. Tukey as 'philosopher.'" *The Annals of Statistics, 2002, 30*, 1619–1228) provides an accessible discussion of Tukey's contributions to the foundations of EDA and related statistical matters.

An insightful treatment of a broad range of issues spanning half a century in the field of data analysis, by a prominent statistical theorist and data analyst, is Peter Huber's *Data Analysis: What Can Be Learned From the Past 50 Years* (Hoboken, NJ: Wiley, 2011).

John Behrens and Chong Ho Yu provide an informative overview of EDA as it applies to psychology (see their "Exploratory Data Analysis" in J. A. Schinka & W. F. Velicer, (Eds., *Handbook of Psychology*, Vol. 2, pp. 33–64. New York,

NY: Wiley, 2003). Their chapter deals with some philosophical aspects of the approach.

Two articles that consider EDA from a Bayesian perspective are I. J. Good, "The Philosophy of Exploratory Data Analysis" (*Philosophy of Science, 50*, 283–295, 1983), and Andrew Gelman, "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing" (*International Statistical Review, 2*, 369–382, 2003).

A highly accessible account of the bootstrap is provided by Julian Simon and Peter Bruce in their "Resampling: A Tool for Everyday Statistical Work" (*Chance, 4*, 22–32, 1991). Simon's book on the same topic, *Resampling: The New Statistics* (Arlington, VA: Resampling Stats, 1997), employs a computer language, RESAMPLING STATS, to simulate the resampling trials.

Chong Ho Yu ("Resampling: A Conceptual and Procedural Introduction," in J. W. Osborne, Ed., *Best Practices in Quantitative Methods*, pp. 283–298. Thousand Oaks, CA: Sage, 2008) provides a useful overview of resampling methodology with an emphasis on its conceptual nature.

The philosopher Jan Sprenger notes in his "Science Without Parametric Models: The Case of Bootstrap Resampling" (*Synthese, 180*, 65–76, 2011) that bootstrap resampling techniques have been ignored by philosophers. He contrasts bootstrap resampling with standard parametric statistical modeling and suggests that the possibilities for fruitfully combining them should be explored.

References

Aiken, L. S., West, S. G., Sechrest, L. B., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: An international survey. *American Psychologist*, *45*, 721–734.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*, 131–160.

Behrens, J. T., Dicerbo, K. E., Yel, N., & Levy, R. (2013). Exploratory data analysis. In J. A. Schinka, W. F. Velicer, and I. B. Weiner (Eds.), *Handbook of psychology* (2nd ed., Vol. 2, pp. 34–70). Hoboken, NJ: Wiley.

Behrens, J. T., & Yu, C-H. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology* (Vol. 1, pp. 33–64). New York, NY: Wiley.

Chatfield, C. (1985). The initial examination of data. *Journal of the Royal Statistical Society*, *Series A*, *148*, 214–254 (with discussion).

Chen, C., Härdle, W., & Unwin, A. (2008). *Handbook of data visualization*. Berlin, Germany: Springer-Verlag.

Dempster, A. P. (2003). John W. Tukey as "philosopher." *The Annals of Statistics*, *30*, 1619–1628.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, *248*, 116–131.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York, NY: Chapman & Hall.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, *13*, 755–779.

Goldman, A. I. (1986). *Epistemology and cognition*. Cambridge, MA: Harvard University Press.

Good, I. J. (1983). The philosophy of exploratory data analysis. *Philosophy of Science*, *50*, 283–295.

Gould, S. J. (1994). The evolution of life on earth. *Scientific American*, *271*, 85–191.

Haig, B. D (2005). An abductive theory of scientific method. *Psychological Methods*, *10*, 371–388.

Haig, B. D. (2013). Detecting psychological phenomena: Taking bottom-up research seriously. *American Journal of Psychology*, *126*, 135–153.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge, MA: MIT Press.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. New York, NY: Springer-Verlag.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York, NY: Wiley.

Huber, P. (2011). *Data analysis: What can be learned from the last 50 years*. Hoboken, NJ: Wiley.

Jones, L. V. (Ed.). (1986). *The collected works of John W. Tukey, Vols. 3 & 4: Philosophy and principles of data analysis.* Monterey, CA: Wadsworth & Brooks/Cole.

Josephson J. R., & Josephson, S. G. (1994). *Abductive inference: Computation, philosophy, technology.* New York, NY: Cambridge University Press.

Lane, D. M., & Sándor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods*, *14*, 239–257.

Laudan, L. (1981). *Science and hypothesis: Historical essays on scientific methodology*. Dordrecht, the Netherlands: Reidel.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science*, *57*, 69–91.

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *American Statistician*, *47*, 217–228.

Lunneborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin*, *98*, 209–215.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Mosteller, F. (1971). The jackknife. *Review of the International Statistical Institute*, *39*, 363–368.

Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, *20*, 355–375.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, *34*, 441–456.

Skinner, B. F. (1984). Methods and theories in the experimental analysis of behavior. *Behavioral and Brain Sciences*, *7*, 511–546.

Sprenger, J. (2011). Science without (parametric) models: The case of bootstrap resampling. *Synthese*, *180*, 65–76.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Thompson, P. A. (1991). Resampling approaches to complex psychological experiments. *Multivariate Behavioral Research*, *26*, 737–763.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614–623.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*, 1–67.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.

Tukey, J. W. (1972). Data analysis, computation, and mathematics. *Quarterly Journal of Applied Mathematics*, *30*, 51–65.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician*, *34*, 23–25.

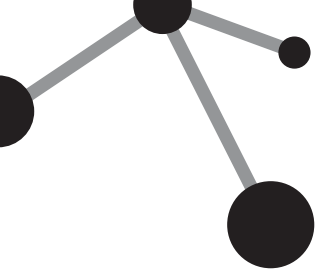
Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*, 393–472.

Yu, C. H. (2003). Resampling methods: Concepts, applications, and justification. *Practical Assessment*, *Research and Evaluation*, *8*, 19 pp.

Yu, C. H. (2008). Resampling: A conceptual and procedural introduction. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 283–298). Los Angeles, CA: Sage.

Yu, C. H. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research*, *3*, 9–22.



3

TESTS OF STATISTICAL SIGNIFICANCE

Although mixing aspects from [Neyman-Pearsonian] and Fisherian tests is often charged as being guilty of an inconsistent hybrid . . . , the error statistical umbrella, linked by the notion of severity, allows for a coherent blending of elements from both approaches.

—D. G. Mayo and A. Spanos, 2011

P-*values should be retained for a limited role as part of the machinery of error-statistical approaches. Even within that system, they need to be supplemented by other devices.*

—S. Senn, 2016

Introduction

It is well known that tests of statistical significance (ToSS) are the most widely used means for evaluating hypotheses and theories in psychology. ToSS have been highly popular in psychology for more than 50 years and in the field of statistics for nearly 80 years. Since the 1960s, a massive critical literature has developed in psychology, and the behavioral sciences more generally, regarding the worth of ToSS (e.g., Harlow, Mulaik, & Steiger, 1997; Hubbard, 2016;

Morrison & Henkel, 1970; Nickerson, 2000). Despite the plethora of critiques of ToSS, most psychologists understand them poorly, frequently use them inappropriately, and pay little attention to the controversy they have generated.

The significance testing controversy shows no signs of abating. Calls for replacing ToSS with alternative statistical methods have been prominent in recent debates. For example, an increasing number of methodologists have expressed a strong preference for the use of Bayesian statistics in place of the most popular form of ToSS, commonly known as *null hypothesis significance testing* (NHST) (e.g., Dienes, 2011; Kruscke, 2015; Wagenmakers, 2007). Also, the so-called new statistics of effect sizes, confidence intervals, and meta-analysis has been assiduously promoted as a worthy package to replace NHST (Cumming, 2014). Some journal editors also have played their part by endorsing alternatives to NHST. For instance, a recent editor of *Psychological Science* endorsed the use of the new statistics wherever appropriate (Eich, 2014), and the current editors of *Basic and Applied Social Psychology* have banned the use of NHST in articles published in their journal (Trafimow & Marks, 2015).

A noteworthy and surprising feature of these calls to do away with NHST is their failure to consider the sensible option of replacing it with defensible accounts of ToSS. The opponents of NHST seem to believe that arguments criticizing the worth of ToSS in its most indefensible form suffice to cast doubt on ToSS in its entirety. However, this is a clear case of faulty reasoning, known as the *fallacy of the false dichotomy*: reject NHST in favor of an alternative that does not involve ToSS, even though there are viable accounts of ToSS available for use.

A major objective of this chapter is to bring two credible perspectives on ToSS to the attention of psychologists. I suggest that these alternative renditions of ToSS can play a legitimate, if limited, role in the prosecution of psychological research. In what follows, I provide a brief overview of NHST and point out its primary defects. I then provide an outline of the neo-Fisherian account of ToSS, which breaks from Neyman and Pearson's formulation and presents an update on Fisher's original position. The second option for a better understanding of ToSS is contained in the contemporary philosophy of statistics known as the *error-statistical philosophy*. The chapter ends with a list of important lessons learned from

the ongoing debates about ToSS that I believe we should carry forward in our thinking on the topic.

Null Hypothesis Significance Testing: Psychology's Textbook Hybrid

Psychologists tend to assume that there is a single unified theory of ToSS. This assumption is primarily based on treatments of the topic furnished by the writers of statistics textbooks in psychology, who pay little, if any, attention to the work of the founding fathers on the topic. By contrast, it is well known in professional statistical circles that there are two major historical theories of ToSS: Fisherian and Neyman-Pearsonian (e.g., Fisher, 1925; Neyman & Pearson, 1933). The relation between the two is a matter of some dispute. It is often said that Neyman and Pearson initially sought to build and improve on Fisher's theory, but that they subsequently developed their own theory as an alternative to that of Fisher. However, historians and theorists in statistics differ on how this relationship should be understood.

A popular view in statistical circles is that there are a number of fundamental points of difference between the two theories, which can be glossed as follows: Both theories adopt fundamentally different outlooks on the nature of scientific method and statistical inference. Fisher argues that an experiment is performed solely to give the data an opportunity to disprove the null hypothesis; no alternative hypothesis is specified, and the null hypothesis is the hypothesis to be nullified. Because one cannot accept the null hypothesis, no provision is made for a statistical concept of power. Fisher (1955) subscribes to an inductive conception of scientific method and maintains that significance tests were vehicles of inductive reasoning. For their part, Neyman and Pearson added the requirement of the specification of an alternative hypothesis and replaced Fisher's evidential *p* value with the Type I error rate. Type II error was admitted, and explicit provision was made for a formal statistical concept of power. Most fundamentally, Neyman and Pearson maintain that significance tests are rules of inductive behavior, not vehicles for inductive reasoning. This gloss on the two schools of thought should serve as a background to the following discussion of their hybridization.

In the behavioral sciences, the best-known account of the hybridized form of ToSS, NHST, is that of Gigerenzer (1993). Elaborating on a metaphor first suggested by Acree (1978), Gigerenzer employs Freudian language to identify the psychological tensions of those who use NHST. As he sees it, features of the Neyman-Pearsonian approach to hypothesis testing combine to form the superego of the hybrid logic and prescribe what should be done. The ego of the hybrid logic, which enables ToSS to be carried out, is that of Fisher. For Gigerenzer, there is a third component of the hybrid, which comes from neither Fisher nor Neyman and Pearson, but from the Bayesian desire to assign probabilities to hypotheses on the basis of the data. Gigerenzer likens this to the Freudian id because it is censored by the Neyman-Pearson superego and the Fisherian ego.

The nature of the psychologists' amalgam and its tensions can, on this received view, be redescribed thus: To the bare bones of Fisherian logic, the hybrid adds the notion of Type II error (opposed by Fisher) and the associated notion of statistical power (Fisher prefers the related notion of experimental sensitivity), but only at the level of rhetoric (thereby ignoring Neyman and Pearson), while giving a behavioral interpretation of both Type I and Type II error (vigorously opposed by Fisher).

There is, however, a further difference attributed to Fisher and Neyman and Pearson, the conflation of which serves to further characterize the amalgam. The inconsistency involves the equation of Fisher's p values with Neyman and Pearson's Type I error rate, in the ubiquitous expression "$p = \alpha$." However, these are said to be fundamentally different things (e.g., Hubbard, 2004). The p values are measures of evidence, closely tied to the data they summarize, whereas α values are rates of error that apply to the tests being used. Fisher, it is said, thought that error rates had no place in his account of significance testing. For their part, Neyman and Pearson are portrayed as thinking that p values had no place in their conception of hypothesis testing. However, the claim that the amalgam brings together two ideas that their originators thought were irreconcilable is challenged by the error-statistical perspective, as I note further in the chapter.

As just seen, Gigerenzer employs the psychodynamic metaphor as a device for organizing some of the sources of confusion that he thinks comprise the hybrid in the minds of many psychological researchers, journal editors, and textbook writers. However, like all

metaphors, it has its limitations. For one thing, it provides a psychological construal of methodological ideas and their relations that might be more illuminatingly cast in more direct methodological terms. For another, it provides a set of hypotheses about the mindset (the "psychic structure") of researchers who employ NHST that lacks proper empirical confirmation. Evidence from protocol analyses of verbal reports of researchers would be required for such confirmation. In addition, this psychological characterization of psychologists' understanding of the hybrid does not take account of the fact that the confusions contained in the amalgam are exacerbated by a tendency of psychologists to misrepresent further the key features of ToSS in a number of ways. For example, levels of statistical significance are taken as measures of confidence in research hypotheses, information about likelihoods is taken as a gauge of the credibility of the hypotheses under test, and reported levels of significance are taken as measures of the replicability of the findings (e.g., Hubbard, 2016). Additional misunderstandings such as these make a psychological characterization of the hybrid beyond the resources of the Freudian metaphor to provide.

It should be said further that there is not a single agreed-upon characterization of the hybrid NHST, as seems to be supposed in treatments of the topic. Halpin and Stam (2006) examined the formulation of the hybrid in six statistics textbooks in psychology published in the period 1940–1960 and found that it received different characterizations. For example, the textbooks differed in the extent to which they made use of ideas from Neyman and Pearson. Relatedly, the authors discovered that the textbooks took ideas from both Fisher and Neyman and Pearson, but that the journal literature that they reviewed made virtually no use of Neyman and Pearson's ideas.

As just intimated, the view that NHST is an inchoate amalgam of Fisher's and Neyman and Pearson's schools of thought is based on the commonly held belief that the two schools are fundamentally different and irreconcilable. However, this belief is not held universally among professional statisticians. For example, Lehmann (1993), a former student of Neyman, maintains that although there are some important philosophical differences between the two schools, the strongly voiced differences of opinion between their founders give the misleading impression that the schools are incompatible. Lehmann contends that, at a practical level, the two approaches are complementary, and that

"*p* values, fixed-level significance statements, conditioning, and power considerations can be combined into a unified approach" (p. 1248). Spanos also adopts the view that the two approaches are complementary. In his well-known textbook (Spanos, 1999), he concludes that the Neyman-Pearsonian approach is suited for testing within the boundaries of a postulated model, whereas the Fisherian approach is suited for testing outside the boundaries of the model. As will be seen, the error-statistical philosophy demonstrates that a number of elements of both of schools of thought can be incorporated in a wide-ranging, coherent position. However, before presenting and discussing the main features of that philosophy, I consider the more circumscribed neo-Fisherian outlook on ToSS.

The Neo-Fisherian Perspective

As its name implies, the neo-Fisherian perspective on ToSS is a reformulation of Fisher's original position. Advocates of this perspective include Cox (2006), Hurlbert and Lombardi (2009), Pace and Salvan (1997), and, to some extent in his later years, Fisher himself. In an extensive recent critical review, Hurlbert and Lombardi (2009) comprehensively surveyed the literature on ToSS and recommend a shift in focus from the original "paleo-Fisherian" and Neyman-Pearsonian classical frameworks to what they maintain is a more defensible neo-Fisherian alternative. For ease of exposition, and convenient reference for the reader, I largely follow the authors' characterization of the neo-Fisherian position. I briefly identify its major elements and indicate how the authors depart from, and see themselves rejecting, the psychologists' hybrid, while improving on problematic elements of Fisher's original position, and rejecting the Neyman-Pearsonian outlook. That said, Hurlbert and Lombardi in fact retain some elements of the latter position, namely alternative hypotheses, power, and confidence intervals.

1. Type I error rate is not specified. In a clear departure from standard practice, critical α's, or probabilities of Type I error, are not specified. Instead, exact *p* values are reported. The publication of Fisher's statistical tables with fixed *p* values was a matter of pragmatic convenience and should not be taken to imply that ToSS requires fixed *p* values to be chosen. Moreover, the refusal to

accept the null hypothesis when an obtained *p* value barely exceeds the adopted value is both rigid and unsound. An α value of .051 has the same evidential import as one of .049.

2. The p values are not misleadingly described as "significant" or "nonsignificant." There is no requirement that the dichotomous "significant"/"nonsignificant" language and thinking be used. Indeed, it is recommended that talk of statistically significant and statistically nonsignificant results be dropped. Undoubtedly, Fisher's publication of critical values of test statistics played a major role in the widespread adoption of this misleading language.

3. Judgment is suspended about accepting the null hypothesis on the basis of high p values. It is not uncommon for textbook authors, and researchers especially, to think that when a *p* value is greater than a specified level of significance, one should accept the null hypothesis as true. However, the neo-Fisherian perspective regards it as neither necessary nor sufficient to accept the null hypothesis on the basis of high *p* values. Factors, such as the strength of experimental conditions, the magnitude of an effect, and power considerations, will have a bearing on whether this belief is sound.

4. The "three-valued logic" that gives information about the direction of the effect being tested is adopted. The logical structure of standard ToSS is a "two-valued logic" by which one chooses between two mutually exclusive hypotheses about the direction of an effect. However, Kaiser (1960), Harris, (1997), and others reason that the researcher who adopts the traditional two-tailed test cannot reach a conclusion about the direction of the effect being tested, and one who employs a one-tailed test cannot conclude that the predicted sign of the effect is wrong. Their proposed solution is to adopt a more nuanced "three-valued logic," where a test for just two hypotheses is replaced by a test of three hypotheses that allows for conclusions about effects with either sign, or an expression of doubt and reserved judgment.

5. Adjunct information about effect sizes and confidence intervals is provided, if appropriate. It is a common criticism of traditional ToSS to decry the overemphasis on *p* values by researchers and their associated neglect of effect sizes and confidence intervals. As noted previously, some methodologists recommend the abandonment of *p*-value statistics in favor of statistics such as these. However, the neo-Fisherian position retains the emphasis on *p* values in significance assessments and regards effect sizes and confidence intervals

as complements to such tests, rather than as alternatives to them. It is important to remember that effect sizes and confidence intervals are faced with their own challenges. For example, the common practice of reporting effect sizes as "small," "medium," and "large" without interpreting them substantively is of limited value. Also, confidence intervals are vulnerable to some of the same charges that are leveled against *p* values, such as the large *n* problem. This problem arises from the fact that discrepancies from any (simple) null hypothesis, however small, can be detected by a (frequentist) ToSS with a large enough sample size (Spanos, 2014).

6. A clear distinction is made between statistical and substantive significance. A source of much confusion in the use and interpretation of ToSS is the conflation of statistical and substantive hypotheses (e.g., Bolles, 1962; Cox, 1958). In the domain of statistical concepts that draws selectively from Fisher and Neyman and Pearson, both the null and the alternative hypotheses are statistical hypotheses. Researchers and textbook writers correctly assume that rejection of the null implies acceptance of the alternative hypothesis, but they also often err in treating the alternative hypothesis as a research, or scientific, hypothesis rather than as a statistical hypothesis. Substantive knowledge of the domain in question is required to formulate a scientific hypothesis that corresponds to the alternative hypothesis. The neo-Fisherian perspective is directly concerned with testing statistical hypotheses as distinct from scientific hypotheses, and it forbids concluding that statistical significance implies substantive significance. At the same time, it urges researchers to explicitly specify the link between the two, warning that sometimes the former may have a small role in establishing the latter.

The neo-Fisherian paradigm contains a package of pragmatic reforms that overcomes some of the problems of NHST, and it improves on aspects of Fisher's original perspective in some respects. Importantly, it represents a reasoned case for retaining *p*-valued significance testing without the focus on hybrid NHST. Although the neo-Fisherian position shares with the error-statistical approach a distrust of the Bayesian outlook on statistics, it differs from the error-statistical approach in rejecting the Neyman-Pearsonian perspective. However, Hurlbert and Lombardi's (2009) claim that the neo-Fisherian position signals the "final collapse" of the Neyman-Pearsonian framework is questionable, for two reasons: First, as noted previously, some elements of Neyman and

Pearson's outlook are retained by the authors. Second, the founder of the error-statistical approach, Deborah Mayo, maintains that the neo-Fisherian approach does not go far enough (reported in Hurlbert & Lombardi, 2009, p. 326), presumably because of its inability to draw key insights from Neyman and Pearson's outlook, such as the notion of error probabilities. In any case, it will become clear that the error-statistical approach provides a more comprehensive outlook on statistical inference than the neo-Fisherian position does.

The Error-Statistical Perspective

An important part of scientific research involves processes of detecting, correcting, and controlling for error, and mathematical statistics is one branch of methodology that helps scientists do this. In recognition of this fact, the philosopher of statistics and science, Deborah Mayo (e.g., Mayo, 1996), in collaboration with the econometrician Aris Spanos (e.g., Mayo & Spanos, 2010, 2011), has systematically developed, and argued in favor of, an *error-statistical* philosophy for understanding experimental reasoning in science. Importantly, this philosophy permits, indeed encourages, the local use of ToSS, among other methods, to manage error.

In the error-statistical philosophy, the idea of an experiment is understood broadly to include controlled experiments, observational studies, and even thought experiments. What matters in all of these types of inquiry is that a planned study permits one to mount reliable arguments from error. By using statistics, the researcher is able to model "what it would be like to control, manipulate, and change in situations where we cannot literally" do so (Mayo, 1996, p. 459). Further, although the error-statistical approach has broad application within science, it is not concerned with all of science, or with error generally. Instead, it focuses on scientific *experimentation* and error *probabilities*, which ground knowledge obtained from the use of statistical methods.

Development of the Error-Statistical Philosophy

In her initial formulation of the error-statistical philosophy, Mayo (1996) modified, and built upon, the classical Neyman-Pearsonian approach to ToSS. However, in later publications with Spanos (e.g.,

Mayo & Spanos, 2011), and in writings with David Cox (Cox & Mayo, 2010; Mayo & Cox, 2010), her error-statistical approach has come to represent a coherent blend of many elements, including both Neyman-Pearsonian and Fisherian thinking. For Fisher, reasoning about *p* values is based on *post-data,* or after-trial, consideration of probabilities, whereas Neyman and Pearson's Type I and Type II errors are based on *pre-data,* or before-trial, error probabilities. The error-statistical approach assigns each a proper role that serves as an important complement to the other (Mayo & Spanos, 2011; Spanos, 2010). Thus, the error-statistical approach partially resurrects and combines, in a coherent way, elements of two perspectives that have been widely considered to be incompatible. In the post-data element of this union, reasoning takes the form of severe testing, a notion to which I now turn.

The Severity Principle

Central to the error-statistical approach is the notion of a severe test, which is a means of gaining knowledge of experimental effects. An adequate test of an experimental claim must be a severe test in the sense that relevant data must be good evidence for a hypothesis. Thus, according to the error-statistical perspective, a sufficiently severe test should conform to the *severity principle*, which has two variants: a *weak severity principle* and a *full severity principle*. The weak severity principle acknowledges situations where we should deny that data are evidence for a hypothesis. Adhering to this principle discharges the investigator's responsibility to identify and eliminate situations where an agreement between data and hypothesis occurs when the hypothesis is false. Mayo and Spanos state the principle as follows:

> Data $\mathbf{x}_0$ (produced by process *G*) do not provide good evidence for hypothesis *H* if $\mathbf{x}_0$ results from a test procedure with a very low probability or capacity of having uncovered the falsity of *H*, even if *H* is incorrect. (Mayo & Spanos, 2011, p. 162)

However, this negative conception of evidence, although important, is not sufficient; it needs to be conjoined with the positive conception of evidence to be found in the full severity principle. Mayo and Spanos formulate the principle thus:

> Data $\mathbf{x}_0$ (produced by process G) provide good evidence for hypothesis H (just) to the extent that test T has severely passed H with $\mathbf{x}_0$. (Mayo & Spanos, 2011, p. 162)

With a severely tested hypothesis, the probability is low that the test procedure would pass muster if the hypothesis was false. Further, the probability that the data agree with the alternative hypothesis must be very low. The full severity principle is the key to the error-statistical account of evidence and provides the core of the rationale for the use of error-statistical methods. The error probabilities afforded by these methods provide a measure of how frequently the methods can discriminate between alternative hypotheses and how reliably they can detect errors.

Error-Statistical Methods

The error-statistical approach constitutes an inductive approach to scientific inquiry. However, unlike favored inductive methods that emphasize the broad logical nature of inductive reasoning (notably, the standard hypothetico-deductive method and the Bayesian approach to scientific inference), the error-statistical approach furnishes context-dependent, local accounts of statistical reasoning. It seeks to rectify the troubled foundations of Fisher's account of inductive inference, makes selective use of Neyman and Pearson's behaviorist conception of inductive behavior, and endorses Charles Peirce's (1931–1958) view that inductive inference is justified pragmatically in terms of self-correcting inductive methods.

The error-statistical approach employs a wide variety of error-statistical methods to link experimental data to theoretical hypotheses. These include the panoply of standard frequentist statistics that use error probabilities assigned on the basis of the relative frequencies of errors in repeated sampling, such as ToSS and confidence interval estimation, which are used to collect, model, and interpret data. They also include computer-intensive resampling methods, such as the bootstrap, Monte Carlo simulations, nonparametric methods, and "noninferential" methods for exploratory data analysis. In all of this, ToSS have a minor, though useful, role.

A Hierarchy of Models

In the early 1960s, Patrick Suppes (1962) suggested that science employs a hierarchy of models that ranges from experimental experience to theory. He claimed that theoretical models, which are high on the hierarchy, are not compared directly with empirical data, which are low on the hierarchy. Rather, they are compared with models of the data, which are higher than data on the hierarchy. The error-statistical approach similarly adopts a framework in which three different types of models are interconnected and serve to structure error-statistical inquiry: primary models, experimental models, and data models. Primary models break down a research question into a set of local hypotheses that can be investigated using reliable methods. Experimental models structure the particular models at hand and serve to link primary models to data models. And, data models generate and model raw data, as well as check whether the data satisfy the assumptions of the experimental models. The error-statistical approach (Mayo & Spanos, 2010) has also been extended to primary models and theories of a more global nature. The hierarchy of models employed in the error-statistical perspective exhibits a structure similar to the important three-fold distinction between data, phenomena, and theory (Woodward, 1989; see also Haig, 2014). These similar three-fold distinctions accord better with scientific practice than the ubiquitous coarse-grained data-theory/model distinction.

Error-Statistical Philosophy and Falsificationism

The error-statistical approach shares a number of features with Karl Popper's (1959) falsificationist theory of science. Both stress the importance of identifying and correcting errors for the growth of scientific knowledge; both focus on the importance of hypothesis testing in science; and both emphasize the importance of strong tests of hypotheses. However, the error-statistical approach differs from Popper's theory in a number of respects: It focuses on *statistical* error and its role in *experimentation*, neither of which were considered by Popper. It employs a range of statistical methods to test for error. And, in contrast with Popper, who deemed deductive inference to be the only legitimate form of inference, it stresses the importance of inductive reasoning in its conception of science.

This error-statistical stance regarding Popper can be construed as a constructive interpretation of Fisher's oft-cited remark that the null hypothesis is never proved, only possibly disproved.

Error-Statistical Philosophy and Bayesianism

The error-statistical philosophy is arguably the major alternative to the reigning Bayesian philosophy of statistical inference. Indeed, in her first major presentation of the error-statistical outlook, Mayo often used Bayesian ideas as a foil in its explication (Mayo, 1996). For one thing, the error-statistical approach rejects the Bayesian insistence on characterizing the evidential relation between hypothesis and evidence in a universal and logical manner in terms of Bayes's theorem via conditional probabilities. It chooses instead to formulate the relation in terms of the substantive and specific nature of the hypothesis and the evidence with regard to their origin, modeling, and analysis. This is a consequence of a commitment to a contextual approach to testing using the most appropriate methods available. Further, the error-statistical philosophy rejects the classical Bayesian commitment to the subjective nature of fathoming prior probabilities in favor of the more objective process of establishing error probabilities understood in frequentist terms. It also finds the turn to "objective" Bayesianism unsatisfactory, but it is not my purpose in this chapter to rehearse those arguments against that form of Bayesianism. Finally, the error-statistical outlook employs probabilities to measure how effectively *methods* facilitate the detection of error and how those methods enable us to choose between alternative hypotheses. Bayesians are not concerned with error probabilities at all. Instead, they use probabilities to measure *belief* in hypotheses or degrees of confirmation. This is a major point of difference between the two philosophies.

Virtues of the Error-Statistical Approach

The error-statistical approach has a number of strengths, which I enumerate at this point without justification: (a) It boasts a philosophy of statistical inference, which provides guidance for thinking about, and constructively using, common statistical methods, including ToSS, for the conduct of scientific experimentation. Statistical methods are often employed with a shallow

understanding that comes from ignoring their accompanying theory and philosophy. (b) It has the conceptual and methodological resources to enable one to avoid the common misunderstandings of ToSS, which afflict so much empirical research in the behavioral sciences. (c) It provides a challenging critique of, and alternative to, the Bayesian way of thinking in both statistics and current philosophy of science; moreover, it is arguably the major modern alternative to the Bayesian philosophy of statistics. (d) Finally, the error-statistical approach is not just a philosophy of statistics concerned with the growth of experimental knowledge. It is also regarded by Mayo and Spanos as a general philosophy of science. As such, its authors employ error-statistical thinking to cast light on vexed philosophical problems to do with scientific inference, modeling, theory testing, explanation, and the like. A critical evaluation by prominent philosophers of science of the early extension of the error-statistical philosophy to the philosophy of science more generally can be found in Mayo and Spanos (2010).

As just noted, the error-statistical perspective addresses a wide range of misunderstandings of ToSS and criticisms of error-statistical methods more generally. Mayo and Spanos (2011) address a baker's dozen of these challenges and show how their error-statistical outlook on statistics corrects the misunderstandings, and counters the criticisms, of ToSS. These include the allegation that error-statistical methods preclude the use of background knowledge, the contention that the fallacies of rejection and acceptance are perpetuated by ToSS, the claim that confidence interval estimation should replace ToSS, and the charge that testing model assumptions amounts to unwarranted data mining. Mayo and Spanos's (2011) reply to these challenges constitutes an important part of the justification of the error-statistical perspective. Because of space limitations, I briefly consider the claims about the fallacies of acceptance and rejection only.

Fallacies of rejection involve the misinterpretation of statistically significant differences. The best known example of such a fallacy is the conflation of statistical and substantive significance, which was discussed previously. This conflation is frequently made by psychological researchers when they employ ToSS. The misinterpretation involves accepting the correctness of a substantive hypothesis solely on the basis of confirming a statistical hypothesis.

This is more likely to happen with a Fisherian use of statistical tests because it carries with it no rival statistical hypothesis to compare with the null hypothesis. Of course, the provision of a statistical alternative to the null, in the manner of Neyman and Pearson, might help to put a brake on those who would otherwise commit the fallacy. The error-statistical perspective incorporates this feature of Neyman and Pearson's approach, explicitly stresses the importance of the distinction between statistical and substantive hypotheses, and urges that it be respected when reasoning back and forth between the data, experimental, and primary models described previously.

Fallacies of acceptance involve taking statistically insignificant differences as grounds for believing that the null hypothesis is true. The basic mistake here is to think that an absence of evidence against the null hypothesis can be taken as evidence for the null hypothesis, as for example when the test used has insufficient power to detect the existing discrepancies. Crucially, the error-statistical approach appeals to the strategy of severe testing to guard against the fallacies of acceptance and rejection. It does this by using post-data assessments of evidence based on the reasoning involved in severe testing. The severity involved formalizes the intuition that *p* values have different evidential import, depending on the size of the sample, or, more generally, the power of the test under consideration (see Mayo & Spanos, 2006, 2011 for details).

What Should We Think About Tests of Significance?

Before concluding this chapter, I enumerate some of the important lessons that I believe can be taken from the extensive debates about the nature and merits of ToSS. Some of these draw from the statistics literature, others from scientific methodology, more generally. These are necessarily presented in brief form. Not all of the material relevant to these lessons has been canvassed in the body of the chapter, but I summon up the chutzpah to present them nonetheless.

1. NHST should not be employed in research. NHST, understood as the variable, inchoate amalgam of elements of Fisherian and Neyman-Pearsonian thinking, should be abandoned because of its incoherence. Its presence in textbooks and research publications

has done, and continues to do, untold damage to psychology. The reasoning in research articles that appeals to the illogic of NHST is either impossible to fathom or the conclusions it gives rise to are unjustified. Psychology's defective statistics education has provided a shallow understanding of ToSS that has resulted in its researchers mechanically employing the hybrid NHST without sufficient awareness of its origins and problems. Moreover, psychology has remained blind to the possibilities of combining elements of different schools of statistical thought in defensible hybrid packages.

2. Defensible forms of ToSS should be employed, where appropriate. It is a mistake to believe that we should give up, or ban, ToSS because of the unsatisfactory nature of its most popular form, NHST. Psychologists are almost entirely unaware that there are credible forms of ToSS, primary among which are the neo-Fisherian and the error-statistical perspectives. Unfortunately, psychology has yet to show an awareness of the fact that these are viable replacements for NHST that can do useful work in data analysis and scientific inference. Methodologists in psychology have a duty to inform themselves about these alternatives to NHST and make considered recommendations about them for researchers in the field. Relatedly, advocates of alternatives to NHST, including some Bayesians (e.g., Wagenmakers, 2007) and the new statisticians (e.g., Cumming, 2014), have had an easy time of it by pointing out the flaws in NHST and showing how their preferred approach does better. However, I think it is incumbent on them to consider plausible versions of ToSS, such as the neo-Fisherian and error-statistical approaches, when arguing for the superiority of their own positions.

3. There are a number of legitimate research goals for ToSS. More specifically, ToSS can do useful local work in different research contexts that involves separating signal from noise. These include pattern detection in exploratory contexts (recommended by Fisher), assistance in judgments about the presence of experimental effects (again, recommended by Fisher [though frequently misused by scientists]), and strong probes designed to detect error in hypotheses under test (a key feature of the error-statistical perspective). Seldom will it be appropriate to rely on p values exclusively (Senn, 2001). Rather, it will mostly be appropriate to employ effect sizes and confidence intervals as complements to ToSS, but that also will depend on context. Generally speaking, I maintain

that these supplements should not be used as replacements for ToSS. Finally, the claim made by some opponents of ToSS that such tests are seldom used in the physical sciences (e.g., McCloskey & Ziliak, 1996) is false (Hoover & Siegler, 2008). ToSS have been, and continue to be, used to good purpose by many researchers in the physical sciences. An instructive example of their informed and rigorous use in physics is the recent discovery of a Higgs boson (van Dyk, 2014).

4. Maintaining the distinction between statistical and substantive hypotheses is of paramount importance. As noted previously, both the neo-Fisherian and the error-statistical perspectives stress the importance of distinguishing between statistical and substantive hypotheses. Despite the fact that ToSS assess statistical hypotheses only, psychologists frequently take them to have direct implications for substantive hypotheses. Moreover, statistical hypotheses play a subservient role to substantive hypotheses and theories, which are the major focus of scientific attention. This is one of a number of reasons why ToSS should have a lesser role to play in the assessment of scientific hypotheses and theories than psychology has generally accorded them.

5. An attitude of strong methodological pluralism should be adopted. The totalizing tendency to be found among some Bayesian statisticians (e.g., Lindley, 2000) and advocates of the Bayesian way in psychology, who argue for the uptake of Bayesian rationality across the board (e.g., Dienes, 2011), should be resisted. The local use of statistics that are fit for purpose is much to be preferred. Similarly, the suggestion of the new statisticians that data analysts should, wherever possible, seek parameter estimates for effect sizes and confidence intervals underappreciates the need for a strong methodological pluralism in which a host of quite different research goals are pursued by employing different statistical methods. Psychology stands to benefit from greater use of additional statistical methods, such as exploratory data analysis, computer-intensive resampling methods, and robust statistics, to mention only a few.

6. Statistical pragmatism is a viable stance. Arguably, an attitude of statistical pragmatism should be encouraged in our use of statistics. Thus, a blending of insights from seemingly opposed schools of statistical thought, which has been built on different philosophical outlooks, is both possible, and sometimes desirable, at the level of practice. For example, thoughtful Bayesian/frequentist

compromises that exploit the insights of both statistical traditions are common in contemporary statistics and some sciences, though they are absent from psychology. Andrew Gelman's heterodox view of Bayesian statistics (e.g., Gelman & Shalizi, 2013) is a good example of the statistical pragmatism I have in mind: It involves the contextual use of Bayesian statistics without buying into the usual inductive Bayesian philosophy of science. Instead, it involves something like a Popperian hypothetico-deductive testing of models, which, moreover, Gelman thinks is consistent with the error-statistical philosophy. This is an example of a "principled" form of pragmatism, in the sense that it comprises an explicitly thought-out philosophy of statistics.

7. Adopting a broad perspective on statistics is important. A broad perspective on statistics is needed to counter the widespread tendency among both scientists and methodologists to view statistics through a narrow lens. Arguably, the error-statistical and Bayesian outlooks are the two most prominent approaches in this regard. The error-statistical approach adopts a broad perspective on the use of statistics in science, as its overview in this chapter makes clear. It has a well-developed philosophy, is concerned with much more than data analysis (e.g., the design of experiments and the validation of model assumptions), and encourages the use of a wide range of statistical methods. The Bayesian outlook on statistics can also be viewed in broad compass, especially if it is joined with a Bayesian philosophy of science and its attendant theory of confirmation—something that most Bayesian statisticians are reluctant do. Further work on the comparative evaluation of the error-statistical and Bayesian perspectives is to be encouraged.

8. There is a need to go beyond standard hypothetico-deductivism in science. The dominant "significant difference" paradigm, with its use of hybridized forms of NHST embedded in an impoverished view of the hypothetico-deductive method, is of questionable value. This paradigm contrasts with the error-statistical perspective and its conception of hypothetico-deductive testing, augmented by a statistical-inductive approach with strong tests. Moreover, hypothesis and theory testing in science is far from all important. Taken together, the tasks of theory construction, including theory generation, theory development, and multicriterial theory appraisal are much more important than just testing for predictive success. One viable replacement for NHST is the *significance sameness* paradigm developed by

Hubbard and Lindsay (e.g., Hubbard, 2016). This paradigm seeks to establish empirical generalizations using effect sizes, confidence intervals, and replication practices, where appropriate, before seeking to understand them through the abductive construction of explanatory theories. Related outlooks on the construction of explanatory theories are to be found in Grice (2011) and Haig (2014).

9. There is a need for different sorts of statistics textbooks. Psychology needs better statistics textbooks, written by specialists who have a good appreciation of modern statistical theory, as well as an understanding of how statistics operate in the prosecution of successful science. To date, statistics textbooks in psychology have been written mainly by nonspecialists, who have made limited use of statistical theory, who have presented NHST as though it were a justified whole, and who have shown a reluctance to replace it with better alternatives. Spanos's *Probability Theory and Statistical Inference* (1999), mentioned previously, is a good example of a textbook that exhibits the desirable features just mentioned. Moreover, his book provides an instructive account of the historical development of ToSS and shows how the Fisherian and Neyman-Pearsonian outlooks can be regarded as complementary. One might expect that its next edition will embrace the fuller-bodied error-statistical outlook.

10. Statistical methods should be taught through methodology. Finally, and importantly, I strongly believe that our understanding of ToSS, and other statistical methods, should be enhanced by greater familiarity with the full range of interdisciplinary contributions to methodology, in addition to our knowledge of statistical practice. Important among these are statistical theory, the philosophy and history of statistics, and statistical cognition. To take just one of these, the value of the philosophy of statistics as an aid to our understanding of ToSS has been considerably underrated by researchers and methodologists in psychology. The error-statistical perspective presented in this chapter is in fact a full-blown philosophy of statistics. As such, it brings with it a deep understanding of the role of ToSS and associated methods, which is made possible by extensive knowledge of the nature of science and its statistical practices, the history and conceptual foundations of statistics, and the philosophy of science more generally (Mayo, 2011, 2012). Philosophy these days is said to be naturalized—that is to say, it is regarded as continuous with science, arguably a *part* of

science, and is concerned with foundational issues *in* science. So located, the philosophy of statistics is well positioned to contribute in important ways to our understanding of statistical theory and practice. Because of this, it deserves to be part of any curriculum that aspires to provide a genuine education in statistics.

Conclusion

Although this chapter is broad-brush in nature, I hope that it will stimulate both psychological researchers and their institutions to think further and deeper about the nature of ToSS and their proper place in research. In more than 50 years of preoccupation with these tests, psychology has concentrated its gaze on teaching, using, and criticizing NHST in its muddled hybrid form. It is high time for the discipline to bring itself up to date with best thinking on the topic and employ sound versions of ToSS in its research.

Further Reading

Jacob Cohen provides a short, but influential, review of the significance testing controversy. See his 'The Earth Is Round ($p < .05$)" (*American Psychologist, 49*, 997–1003, 1994).

Raymond Nickerson provides an excellent extensive review of the ongoing controversy surrounding null hypothesis significance testing in his "Null hypothesis Significance Testing: A Review of an Old and Continuing Controversy" (*Psychological Methods, 5*, 241–301, 2000).

Peter Halpin and Henderikus Stam offer an informative account of the hybridization of Fisher's and Neyman and Pearson's approaches to significance testing in psychology over the period 1940–1960. See their "Inductive Inference or Inductive Behavior: Fisher and Neyman-Pearson Approaches to Statistical Testing in Psychological Research (1940–1960)" (*American Journal of Psychology, 119*, 625–653, 2006).

In his book, *Statistical Inference: A Commentary for the Social and Behavioral Sciences* (New York, NY: Wiley, 1986), Michael Oakes undertakes a critical examination of the different schools of statistical thought and the misuse of null hypothesis significance testing in the social and behavioral sciences.

An accessible philosophical examination of both scientific and statistical inference in science, with particular reference to psychology, is provided by Zoltán Dienes in his book, *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference* (Basingstoke, England: Palgrave, 2008).

Denton Morrison and Raymon Henkel's *The Significance Test Controversy: A Reader* (Piscataway, NJ: Aldine, 1970) is a collection of major articles on the significance testing controversy published in sociology and psychology prior

to 1970. In an essay review of this volume, Ronald Giere provides a useful philosophical discussion of the controversy. See his "The Significance Test Controversy" (*British Journal for the Philosophy of Science, 23*, 170–181, 1971).

An important collection of more recent assessments of the worth of statistical significance tests can be found in the reader, *What If There Were No Significance Tests?* (L. L. Harlow, S. A. Mulaik, and J. H. Steiger, Erlbaum, Mahwah, NJ, 1997).

An excellent informative assessment of ToSS and their place in behavioral science research is provided by Raymond Hubbard in his book *Corrupt Research* (Thousand Oaks, CA: Sage, 2016). Hubbard recommends the adoption of the *significance sameness paradigm*, in which the researcher seeks to establish empirical generalizations using effect sizes, confidence intervals, and replication practices, where appropriate, before seeking to understand them through the abductive construction of explanatory theories.

A defensible approach to Fisherian statistical inference, modified in the face of criticisms of Fisher's classical approach, is provided by Stuart Hurlbert and Celia Lombardi in their article, "Final Collapse of the Neyman-Pearson Decision-Theoretic Framework and Rise of the NeoFisherian" (*Annales Zooologici Fennici, 46*, 311–349, 2009).

The following two books chart the development of the error statistics research program: Deborah Mayo, *Error and the Growth of Experimental Knowledge* (Chicago, IL: University of Chicago Press, 1996); Deborah Mayo and Aris Spanos (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (Cambridge, England: Cambridge University Press, 2010). A shorter, more accessible, account of the error statistics approach is Deborah Mayo and Aris Spanos, "Error Statistics," in Prasanta Bandyopadhyay and Malcolm Forster (Eds.), *Handbook of the Philosophy of Science, Vol. 7: Philosophy of Statistics* (pp. 153–198) (Amsterdam, the Netherlands: Elsevier, 2011).

David Grayson provides an instructive examination of different competing conceptions of probability presupposed in statistical inference and their problems for understanding, and use of, statistical inference. See his "The Frequentist Façade and the Flight From Evidential Inference" (*British Journal of Psychology, 89*, 325–345, 1998).

References

Acree, M. C. (1978). *Theories of statistical inference in psychological research: A historico-critical study.* Ann Arbor, MI: University Microfilms No. H790 H7000.

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, *11*, 639–645.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, *29*, 357–372.

Cox, D. R. (2006). *Principles of statistical inference.* Cambridge, England: Cambridge University Press.

Cox, D. R., & Mayo, D. G. (2010). Objectivity and conditionality in frequentist inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent*

exchanges on experimental reasoning, reliability, and the objectivity and rationality of science (pp. 276–304). New York, NY: Cambridge University Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society*, *Series B*, *17*, 69–78.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8–38.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. San Diego, CA: Academic Press.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences.* Cambridge, MA: MIT Press.

Halpin, P. F., & Stam, H. J. (2006). Inductive inference or inductive behavior: Fisher and Neyman-Pearson approaches to statistical testing in psychological research (1940–1960). *American Journal of Psychology*, *119*, 625–653.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). (1997). *What if there were no significance tests?* (pp. 145–174). Mahwah, NJ: Erlbaum.

Hoover, K. D., & Siegler, M. V. (2008). Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology*, *15*, 1–37.

Hubbard, R. (2004). Alphabet soup: Blurring the distinction between *p*'s and α's in psychological research. *Theory & Psychology*, *14*, 295–327.

Hubbard, R. (2016). *Corrupt research: The case for reconceptualising empirical management and social science*. Los Angeles, CA: Sage.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, *46*, 311–349.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*, 160–167.

Kruscke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam, the Netherlands: Elsevier.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*, 1242–1249.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*, 293–319.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge.* Chicago, IL: University of Chicago Press.

Mayo, D. G. (2011). Statistical science and philosophy of science: Where do/should they meet in 2011 (and beyond)? *Rationality*, *Markets*, *and Morals*, *2*, 79–102.

Mayo, D. G. (2012). Statistical science meets philosophy of science, Part 2: Shallow versus deep explorations. *Rationality*, *Markets*, *and Morals*, *3*, 71–107.

Mayo, D. G., & Cox, D. (2010). Frequentist statistics as a theory of inductive inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 247–304). New York, NY: Cambridge University Press.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, *57*, 323–357.

Mayo, D. G., & Spanos, A. (Eds.). (2010). *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. New York, NY: Cambridge University Press.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of philosophy of science, Vol. 7: Philosophy of statistics* (pp. 153–198). Amsterdam, the Netherlands: Elsevier.

McCloskey, D. N., & Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, *34*, 97–114.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy: A reader*. Chicago, IL: Aldine.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, *231*, 289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Pace, L., & Salvan, A. (1997). *Advanced Series on Statistical Science and Applied Probability, Vol. 4. Principles of statistical inference from a neo-Fisherian perspective*. Singapore: World Scientific.

Peirce, C. S. (1931–1958). *The collected papers of Charles Sanders Peirce, Vols. 1–8*. C. Hartshorne & P. Weiss (Eds., Vol. 1–6) & A. W. Burks (Ed., Vol. 7–8). Cambridge, MA: Harvard University Press.

Popper, K. R. (1959). *The logic of scientific discovery*. London, England: Hutchinson.

Senn, S. (2001). Two cheers for *p*-values? *Journal of Epidemiology and Biostatistics*, *6*, 193–204.

Senn, S. (2016). Are *p*-values the problem? *American Statistician* [Online discussion of The American Statistical Association statement on statistical significance and p-values].

Spanos, A. (1999). *Probability theory and statistical inference: Economic modeling with observational data.* Cambridge, England: Cambridge University Press.

Spanos, A. (2010). On a new philosophy of frequentist inference: Exchanges with David Cox and Deborah G. Mayo. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 315–330). New York, NY: Cambridge University Press.

Spanos, A. (2014). Recurring controversies about *p* values and confidence intervals revisited. *Ecology*, *95*, 645–651.

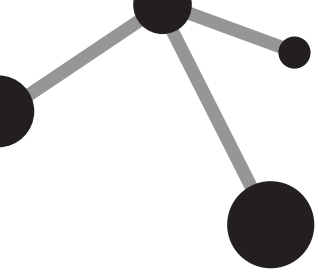
Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 International Congress* (pp. 252–261). Stanford, CA: Stanford University Press.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.

Van Dyk, D. A. (2014). The role of statistics in the discovery of a Higgs boson. *Annual Review of Statistics and Its Applications*, *1*, 41–59.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*, 393–472.



4

BAYESIANISM

[S]cientific reasoning is reasoning in accordance with the calculus of probabilities.

—C. Howson and P. Urbach, 2006, p. 10

[It] is hard to see what motivates the Bayesian who wants to replace the fabric of science . . . with a vastly more complicated representation in which each statement of science is accompanied by its probability, for each of us.

—H. Kyburg, 1992, p. 149

Introduction

Bayesianism is a formal theory of reasoning based on probability theory. It deals with a number of important, and related, general ideas, such as rationality, confirmation, and inductive inference, including statistical inference. Bayesianism takes its name from the fact that it makes central use of a theorem derived from the probability calculus known as *Bayes's theorem*, a theorem that is regarded as a major constraint on how one should rationally modify one's opinions or beliefs in the light of incoming evidence.

Bayesianism boasts a number of different, but related, movements. Within philosophy, these include the overlapping spheres of Bayesian epistemology, philosophy of science, and confirmation theory. Outside philosophy, they include statistics and learning theory. This chapter largely focuses on Bayesian confirmation theory, although some attention is given to the philosophy of Bayesian statistics. Even with this circumscribed focus, my examination of Bayesianism is highly selective. I use the general term *Bayesianism* to refer to these two different strands of Bayesian thinking: Bayesian confirmation theory and Bayesian statistics. Although they have some things in common, they have developed largely independently of each other, for Bayesian confirmation theory has focused on probability rather than statistics. Given their appreciable differences, I discuss them more or less separately, even though their separation is artificial. Howson and Urbach's (2006) popular book on Bayesian reasoning is a rare attempt to provide a unified account of scientific methods that incorporates both aspects of Bayesianism.

The existence of these two major strands of Bayesianism should caution us from thinking that Bayesianism is, or could be, a unique, or even unified, position. In fact, the Bayesian philosopher and statistician I. J. Good (1971) calculated that there are at least 46,656 varieties of Bayesians and noted that there are more distinctive positions to fill than there are Bayesians to fill them. Not only does Bayesianism take many forms, but also it is multifaceted in nature, as the introductory remarks might imply. This diversity underscores just how selective my treatment of Bayesian thinking is. In addition, the treatment is largely conceptual and informal and leaves aside many of the technical aspects of Bayesian thinking.

In what follows, I briefly record psychology's attitudes to Bayesianism. I then trace some of the broad contours of Bayesian confirmation theory. In turn, these involve a consideration of Bayes's theorem, the Bayes factor, the ability of Bayesian confirmation theory to improve on the hypothetico-deductive method, and the question of whether Bayesianism provides an illuminating account of the approach to theory evaluation known as inference to the best explanation. Thereafter, I present an evaluation of the important philosophy of Bayesian statistical practice developed by Gelman and Shalizi (2012), which combines elements of

philosophy of science and Bayesian statistics in a creative manner that has relevance for researchers in psychology. The chapter concludes by bringing together the main points in the form of a summary and offers a few broad recommendations for practice.

Bayesianism in Psychology

There are two striking facts about the presence of Bayesian thinking in psychology. One is that there is virtually no reference in the methodological literature to Bayesian confirmation theory, reflecting, I think, a general lack of appreciation for the value of contemporary philosophy of science for understanding both statistics and science. The other notable fact is that Bayesian statistics has taken an age to assert itself in psychology and related sciences.

Although Bayesian statistical thinking pre-dated the advent of classical statistics, the latter has dominated thinking in the field of statistics since the 1930s. However, Bayesian statistical theory and practice have been steadily developing and increasing their influence in the fields of both statistics and science. The situation is different in psychology. When Edwards, Lindman, and Savage (1963) brought the Bayesian outlook in statistics to psychology over 50 years ago, they acknowledged that it was a new perspective lacking full coherence. In addition, there existed at that time no textbook that made the arsenal of Bayesian methods available to psychological researchers. Matters have steadily improved since then. Major advances have been made in Bayesian statistical theory and practice, a number of accessible textbooks have been written for behavioral and social scientists, position papers and tutorials advocating and expositing Bayesian statistical ideas have been published, and computer programs for implementing a variety of Bayesian methods have been developed. Currently, a cadre of methodologists has advocated the uptake of Bayesian statistical methods in psychology (e.g., Dienes, 2011; Kruscke, 2015; Wagenmakers, 2007), but the enthusiasm of these methodologists has not been matched by the discipline's research fraternity. In all of these developments, neither the philosophy of Bayesian statistics nor Bayesian confirmation theory has been visible.

Similarly, prominent institutional efforts to reform psychology's use of statistical methods essentially ignore the Bayesian alternative. Instead, they continue for the most part to employ classical

statistical methods. For example, the American Psychological Association's Task Force on Statistical Inference (Wilkinson & the Task Force on Statistical Inference, 1999) was charged with considering alternatives to tests of statistical significance but said nothing about the place of Bayesian statistical methods in psychological research. More recently, the Association for Psychological Science's promotion of the "new statistics" (Cumming, 2014; Eich, 2014) urged replacement of null hypothesis significance testing with frequentist confidence intervals, effect sizes, and meta-analysis. Barely a mention was made of Bayesian methods.

Bayesian Confirmation Theory

What is it for empirical evidence to provide confirmation or disconfirmation of a scientific hypothesis or theory? Methodologists of science have worked long and hard to answer this important and challenging question by developing theories of scientific confirmation. Despite the considerable fruits of their labors, there is widespread disagreement about which theory of confirmation we should accept. Over time, a large number of philosophers of science have contributed to Bayesian confirmation theory (e.g., Earman, 1992; Horwich, 1982; Howson & Urbach, 2006; Rosenkrantz, 1977). Many philosophical methodologists now believe that Bayesianism confirmation theory holds the best hope for building a comprehensive and unified theory of scientific inference.

Bayes's Theorem

As noted at the outset, the Bayesian approach to scientific inference is so called because it makes central use of a theorem of the mathematical calculus of probability known as *Bayes's theorem*. The theorem is widely thought to have originated with the nineteenth-century mathematician and clergyman Reverend Thomas Bayes. However, there are two interesting points of historical accuracy that are worth mentioning in this regard. First, there is a small irony in the fact that, in his original essay, Bayes (1764) does not explicitly state Bayes's theorem to solve the problem he addresses (providing a justification for the uniform prior distribution) (Earman, 1992). Second, Bayes himself was likely not the originator of Bayes's theorem. The historian of statistics, Stephen Stigler (1983), estimated a

posterior probability of 3 to 1 that the evidence uncovered thus far favors the Cambridge mathematician Nicholas Saunderson as the discoverer of the theorem.

Bayes's theorem can be expressed in different forms that are used for different purposes. For expository convenience, I present it here in its simplest form, which deals with one hypothesis. The use of Bayes's theorem for testing two hypotheses is discussed further in the chapter. In this simple form, it is written as follows:

$$\Pr\left(H/D\right) = \frac{\Pr\left(H\right) \times \Pr\left(D/H\right)}{\Pr\left(D\right)}$$

With the proviso that Pr(D) and Pr(H) cannot be zero or one (because each would determine by itself the same value for the resulting posterior probability), the theorem says that the posterior probability of the hypothesis H is obtained by multiplying the prior probability of the hypothesis Pr(H) by the probability of the data, given the hypothesis Pr(D/H) (the likelihood), and dividing the product by the prior probability of the data Pr(D). I note in passing that in this chapter T (theory) is sometimes substituted for H, and E (evidence) is sometimes substituted for D.

It is through use of this and other versions of Bayes's theorem that Bayesians are able to implement their view of scientific inference, which is the orderly revision of opinion on the basis of new information. To achieve this goal, Bayesians employ Bayes's theorem iteratively. Having obtained a posterior probability assignment for their hypothesis via Bayes's theorem, they can then go on and use that posterior probability as the new prior probability in a further use of Bayes's theorem designed to yield a revised posterior probability and so on. In this way, the Bayesian inquirer learns from experience.

For Bayesians, a couple of attractive features of this gloss on Bayesian scientific inference are often emphasized. Most important, the Bayesian approach is said to square with the stated purpose of scientific inquiry noted previously, namely, securing the probability of a hypothesis in the light of the relevant evidence. By contrast, the informational output of classical statistical inference in science is the probability of the data, given the truth of the null hypothesis, but it is just one input in the Bayesian scheme of things. A second

stated desirable feature of the Bayesian view is its willingness to make use of relevant information about the hypothesis before the empirical investigation is conducted and new data are obtained, explicitly in the form of a prior probability estimate of the hypothesis. Traditional statistical inference assumes that inferences should be based solely on present data, without any regard for what might be brought to a study in the way of belief or knowledge about the hypothesis to be tested—a position that Bayesians contend is hardly designed to maximize the chances of learning from experience.

There is, in fact, a third positive feature claimed to be associated with Bayes's theorem: From the common situation where different prior probabilities are assigned to a given hypothesis, the accumulated evidence obtained through repeated use of Bayes's theorem will see those discrepant hypotheses converge. Moreover, it is a feature of Bayesian confirmation theory that the higher the posterior probabilities of hypotheses are, the stronger the hypothesis is said to be confirmed, and the more rational it is to accept it or believe in its truth.

Statistical Hypothesis Testing: The Bayes Factor

An important goal of science is to test hypotheses or theories. One recommended Bayesian way of reaching this goal is to calculate so-called Bayes factors. The Bayes factor, sometimes called the *posterior odds ratio*, can be understood in a general way to apply to multiple hypotheses. For convenience, I focus here on the simplest case of two hypotheses. In this context, the Bayes factor can be represented in the following simple equation (as with Bayes's theorem, D can be taken as empirical evidence more generally; H can be construed as a theory; and the theory can be understood as a statistical model):

$$BF_{12} = \frac{\Pr(D/H_1)}{\Pr(D/H_2)}$$

In words, the Bayes factor for two hypotheses, H_1 and H_2, grades the impact of the evidence D on the two hypotheses by comparing the probability of the observed data, given the truth of H_1, with the probability of the observed data, given the truth of H_2. The

Bayes factor is a ratio represented by a number, which quantifies the strength of evidence in favor of one hypothesis over its rival.

Orthodox Bayesians claim three principal advantages for the Bayes factor (e.g., Andraszewicz et al., 2015): The first claimed advantage is that it quantifies the evidence in a precise way. Jeffreys (1961, Appendix B) provides an interpretive guide for different Bayes factors, which, with minor modifications, is generally accepted today. A Bayes factor of less than 1 is taken as negative evidence, 1–3 as anecdotal evidence, 3–10 as moderate evidence, 10–30 as strong evidence, 30–100 as very strong evidence, and more than 100 as extreme evidence. A second claimed advantage for the Bayes factor is that it requires tests to be carried out in a comparative manner because it explicitly weighs the evidence for one hypothesis in relation to another. The strategy of comparative hypothesis evaluation is widely endorsed by scientific methodologists. The third claimed advantage of the Bayes factor is that it is coherent and is not beset with the incoherencies that attach to traditional *p*-value hypothesis testing.

At the same time, orthodox Bayesians acknowledge that there are a number of challenges for their approach to hypothesis testing. The objection most relevant to our present purpose is the claim that hypothesis testing should be replaced by estimation. In reply, Andraszewicz et al. (2015) contend that "hypothesis testing is a legitimate scientific endeavor that requires a proper statistical implementation" (p. 527). They maintain that there are legitimate scientific questions that cannot be addressed by an estimation framework (e.g., Can people anticipate the future? Is there a gene for Alzheimer disease? Does the Higgs boson exist?), but acknowledge that once an effect has been detected, its size can be gauged by means of estimation.

Further in the chapter, I present Gelman and Shalizi's (2012) neo-Popperian philosophy of Bayesian statistics, which regards the Bayes factor (and other Bayesian statistical methods) in a different light from orthodox Bayesians. The gist of their view is that the Bayes factor can be a useful tool for predicting and understanding structure in the data. However, they caution against thinking of models tested by the Bayes factor as true or taking the posterior probabilities of models too seriously. The presentation of their philosophy, which motivates these attitudes, should help to clarify these points about truth and posterior probabilities.

Bayesianism and the Hypothetico-Deductive Method

One of the clear achievements of Bayesianism is its ability to improve on the unsatisfactory approach to hypothesis and theory appraisal taken by the standard hypothetico-deductive method. The hypothetico-deductive method has long been the method of choice for the evaluation of scientific theories (Laudan, 1981), and it continues to have a dominant place in psychology. Despite its popularity, it is usually characterized in an austere manner: The researcher takes a hypothesis or theory of interest and tests it indirectly by deriving from it one or more observational predictions that are themselves directly tested. Predictions borne out by the data are taken to confirm the theory to some degree; those predictions that do not square with the data count as disconfirming instances of the theory. Normally, the theory is not compared with rival theories in respect of the data, only with the data themselves.

The hypothetico-deductive method, in something like this form, has been strongly criticized by methodologists on a number of counts (e.g., Glymour, 1980; Rozeboom, 1997). One major criticism of the method is that it is confirmationally lax. This laxity arises from the fact that any positive confirming instance of a hypothesis submitted to empirical test can confirm any hypothesis that is conjoined with the test hypothesis, regardless of how plausible it might be. This state of affairs is known as the *fallacy of irrelevant conjunction*, or the tacking problem, because confirmation of a test hypothesis also confirms any conjunct that is attached to the test hypothesis. The fallacy of irrelevant conjunction arises with the hypothetico-deductive method because predictions are deduced from hypotheses only by making use of auxiliary hypotheses drawn from background knowledge, and some of the background knowledge drawn on may not be pertinent to the matter at hand.

Clearly, this is an unacceptable state of affairs. Bayesians have challenged the assumption that the occurrence of the consequences of a theory confirm the theory and its conjuncts holistically. They argue that the Bayesian approach enables the differential support of the elements of a theory, specifying conditions showing that evidence never increases the probability of a theory conjoined with any additional hypothesis by more than it increases the probability of that theory.

Another major criticism of the hypothetico-deductive method is that it tests a single hypothesis or theory of interest against the empirical evidence; it does not test a hypothesis or theory in relation to rivals in respect of the evidence. This is held to be a major flaw because it is widely agreed that theory evaluation is a comparative affair involving simultaneous evaluation of two or more hypotheses or theories.

The comparative nature of theory evaluation is straightforwardly handled by the Bayesian position by rewriting the simple form of Bayes's theorem given previously to deal with two or more hypotheses. Here, Bayes's theorem is presented for the case of two hypotheses, where the theorem can be written for each hypothesis in turn. For the first hypothesis,

$$\Pr\left(H_1/D\right) = \frac{\Pr\left(H_1\right) \times \Pr\left(D/H_1\right)}{\Pr\left(H_2\right) \times \Pr\left(D/H_2\right) + \Pr\left(H_1\right) \times \Pr\left(D/H_1\right)}$$

This says that the posterior probability of the first hypothesis is obtained by multiplying its prior probability by the probability of the data, given that hypothesis (the likelihood), and dividing the product by the value that results from adding the prior probability of the second hypothesis, multiplied by the likelihood for that hypothesis, to the prior probability of the first hypothesis, multiplied by its likelihood. Bayes's theorem for the second hypothesis is written in a similar way.

It would seem, then, that the confirmational worth of Bayesianism is superior to that of the standard hypothetico-deductive account of scientific method. However, there are now available more sophisticated accounts of hypothetico-deductive reasoning that do not suffer the defects of the standard view. I briefly refer to the improved outlook on hypothetico-deductive reasoning when I discuss Gelman and Shalizi's (2012) philosophical framework for Bayesian model testing.

Bayesianism and Inference to the Best Explanation

Recently, some Bayesians have claimed that their perspective on scientific method can also provide an enhanced characterization of the important approach to theory evaluation known as

inference to the best explanation. Inference to the best explanation is based on the belief that much of what we know about the world is based on considerations of explanatory worth. In contrast to the Bayesian approach, received accounts of inference to the best explanation take theory evaluation to be a qualitative exercise that focuses on explanatory criteria, not a quantitative undertaking in which one assigns probabilities to theories (Haig, 2009). For example, Paul Thagard's (1992) account of inference to the best explanation, known as the *theory of explanatory coherence*, employs the three criteria of explanatory breadth, simplicity, and analogy, which all directly have to do with explanation.

Although inference to the best explanation has typically been regarded as a competitor for Bayesian theory evaluation, Lipton (2004) argues that the two approaches are broadly compatible, and that, in fact, their proponents "should be friends." In broad terms, he suggests that judgments of the *loveliest* explanation, which are provided by the evaluative criteria of inference to the best explanation, such as unificatory power, precision, and elaboration of explanatory mechanisms, contribute to assessments of the *likeliest* explanation, which are provided by the probabilities of the Bayesian approach. Lipton maintains that the explanatory considerations invoked in inference to the best explanation guide determination of the prior probabilities (and the likelihoods) that are inserted in Bayes's theorem.

However, although appeal to explanatory matters might be one way in which Bayesians can determine their prior probabilities, Lipton does not suggest how this might be done. Further, those who hold inference to the best explanation to be a normative approach to scientific theory evaluation, with its own distinctive character, will worry that Lipton relegates it to a descriptive role within a Bayesian normative framework (e.g., Psillos, 2004).

Another way of showing the compatibility of inference to the best explanation and Bayesianism is to translate the evaluative criteria employed within inference to the best explanation into probabilistic terms. McGrew (2003) has done this by taking the important theoretical virtue of consilience, or explanatory breadth, and showing that its Bayesian form leads to higher posterior probabilities of the hypotheses being evaluated. Nevertheless, McGrew acknowledges that by translating consilience into its "flattened" probabilistic form, it no longer remains a genuine explanatory

virtue: Not only is there no guarantee that consilience will be concerned with an *explanation* of the evidence, there is no way that probabilistic translations of the explanatory virtues can refer to the causal connections that are often appealed to in scientific explanations. Further, Weisberg (2009) argues that the explanatory loss incurred in such translations will occur for any distinctively explanatory virtue that is given such probabilistic treatment.

In short, it would seem that Bayesianism cannot capture the depth of the intuitively important notion of explanatory power. Thus, although qualitative accounts of inference to the best explanation, such as the theory of explanatory coherence, can be clothed in probabilistic dress, they are best used on their own terms or appraising scientific theories (Thagard, 2000).

I turn now to consider two of the most important criticisms that have been leveled at Bayesian confirmation theory: the problem of the priors and the problem of old evidence.

Two Common Criticisms of Bayesianism

The Problem of the Priors

It is often said that the use of Bayes's theorem, and its attendant appeal to prior probabilities, is the most distinctive feature of Bayesianism. Further, it is frequently claimed that a major advantage of employing prior probabilities is that it enables the investigator to explicitly incorporate relevant information into determining knowledge claims in addition to the immediate evidence at hand. However, it is acknowledged by Bayesians and non-Bayesians alike that the estimation and interpretation of prior probabilities presents a number of difficulties. Here, I briefly consider two ways in which Bayesians deal with the interpretation of prior probabilities, arising from the adoption of different understandings of the nature of probability.

The most prominent, and historically influential, strand of Bayesianism takes probabilities to be subjective degrees of belief, and for this reason, it is often called *subjectivist Bayesianism*. Bayesians of this type, such as Leonard Savage (1964), regard probabilities as personal degrees of belief held by individuals. Thus, the value of the prior probability of a hypothesis or theory is the actual degree of belief that an individual cares to assign to the hypothesis or theory.

There are a number of criticisms of this interpretation that point to the arbitrary nature of such probability assignments and the unacceptably high degree of subjectivity involved, both of which follow from scientists using different means, and offering different personal reasons, for arriving at different estimates. The standard subjectivist reply is to point out that individuals successively modify their earlier divergent probability estimates via Bayes's theorem as each set of fresh evidence comes in. Moreover, some subjectivists argue that, under certain conditions, initially discrepant priors eventually converge to the same posterior value (or "wash out") as evidence accumulates in favor of a hypothesis or theory. Thus, subjectivists claim that by reaching consensus in this way, the objectivity in subjectivist Bayesianism is made clear.

In an effort to ensure that reasonable priors are estimated, some subjectivists maintain that we need to add additional principles to Bayes's theorem. Abner Shimony's (1993) *tempered personalism* is a well-known approach to supplementing pure subjective Bayesianism, but here I mention the account of tempered personalism due to Wesley Salmon (1990). In an effort to avoid the strong subjectivity of the subjectivist position, Salmon opts for an objective frequentist interpretation of probability (Salmon (1966, 1970, 1990) in which plausibility assessments are considered important. Salmon takes a number of Kuhn's (1970) well-known criteria for theory choice and uses them as constraints in estimating prior probabilities. He reasons as follows: (a) Because the criterion of simplicity is often construed as an a priori virtue, it can figure in the determination of the prior probabilities of Bayes's theorem; (b) consideration of the criterion of consistency demands that a hypothesis must have a prior probability greater than zero; (c) the criterion of the external consistency of a hypothesis (one that fits well with other accepted hypotheses) should receive a high prior probability; and (d) the criterion of the fruitfulness of a hypothesis can be understood in two ways—as prior probabilities having to do with unification and as the likelihoods in Bayes's theorem, understood in terms of their effectiveness in predicting new phenomena.

Salmon (1990) believes that we assign a prior probability to a new hypothesis to estimate the objective probability that the hypothesis will turn out to be true. Although the use of the criteria just mentioned results in exact values for prior probabilities, the accuracy of their values does not matter to Salmon. For he, like

others, maintains that, in the end, the discrepant values wash out in the face of accumulating evidence.

The Problem of Old Evidence

Scientists and philosophers have often noted that hypotheses and theories receive a boost in confirmation if they predict novel evidence. Reasonably enough, Bayesians are quick to point out that, through the use of Bayes's theorem, their theory of confirmation is readily able to account for this methodological intuition.

By contrast, taking old evidence as support for new theories is often seen as a major problem for Bayesian confirmation theory, whether it be subjectivist or objectivist in form. The problem of old evidence, as it is known, was introduced by Clark Glymour (1980), and is as follows: Scientists sometimes find support for a theory in the form of evidence that was known prior to the introduction of that theory. As one of several examples, Glymour cites the already-known anomalous advance of the perihelion of Mercury being taken as strong support for Einstein's general theory of relativity. Although this sort of confirmational relation is often appealed to in science, Bayesians have great difficulty making sense of it. First, old evidence, as established fact, should be assigned a probability of 1 (or something very close to it). Second, the probability of evidence under any theory or hypothesis should also be 1. Now, if we enter these values into Bayes's theorem, both the posterior and the prior probabilities of the theory in question come out the same. Thus, in cases like these, the old evidence in fact bears no relation to the credibility of the theory and therefore cannot confirm it. It is the failure of Bayesian confirmation theory to deal with this type of relationship between theory and evidence that worries its critics. It should be recalled that Bayes's theorem is coupled with the proviso that the prior probabilities of the data and the hypothesis cannot be 0 or 1 in order to safeguard against cases like this.

A number of philosophers have claimed that this problem of old evidence is unsolvable in principle, and that, for this and perhaps other reasons, Bayesian confirmation theory should be rejected (e.g., Glymour, 1980; Leplin, 1997). However, it should be said that Bayesianism is a resourceful theory of confirmation, and several solutions to the problem of old evidence have been advanced. One solution has been offered by Howson and Urbach (2006). They

defend the use of Bayes's theorem with old evidence by arguing that the background knowledge used to determine the prior probability of the theory should be confined to existing background beliefs and should explicitly exclude knowledge of the old evidence. The authors justify this move by maintaining that the purpose of scientific research is solely to gauge the impact of the existing evidence on the probability of the theory. One worry about this proposal is based on the expectation that confirmation of a theory should be based on the full range of relevant background beliefs. More generally, there is no consensus among Bayesians regarding which solution to the problem of old evidence is best, and criticisms of the various solutions have been fashioned by Bayesian and non-Bayesians alike. I refer the reader to Earman (1992) for a critical assessment of the matter.

What Should We Think About Bayesian Confirmation Theory?

Philosophical assessments of the worth of Bayesian confirmation theory range from claims that it is without peer as a theory of scientific reasoning to the view that it is fundamentally wrong-headed. Howson and Urbach (2006) exemplify the former view, claiming that scientific reasoning is both inductive and probabilistic, and that the axioms of probability suffice to articulate such reasoning. Many scientists and philosophers reject this view on the grounds that it omits important forms of scientific inference, such as abductive reasoning, and, moreover, that a good deal of science involves both nonprobabilistic quantitative inference and qualitative inference as well. The latter view is held by the philosopher of science, Mario Bunge (2008), who argues that Bayesianism is fundamentally wrong for three reasons: It assigns probabilities to statements rather than taking them as objective features of the world; it conceives of probabilities as subjective degrees of belief, which have no scientific standing; and it appeals to probabilities in the absence of the proper requirement of randomness. For these reasons, Bunge judges Bayesianism to be a pseudoscience. This is an extreme judgment and is based on a particular view of probability with which many Bayesians and non-Bayesians will disagree. Bunge's view is mentioned here to illustrate just how negative some assessments of Bayesian confirmation theory can be. Finally, some

advocates of Bayesianism see it as a comprehensive theory of confirmation applicable to all of science, whereas others see it as having only context-specific applications.

A strong criticism of approaches to confirmation in terms of probabilities that should be taken seriously is that of Kelly and Glymour (2004). It continues Glymour's (1980) earlier well-known critique of Bayesian confirmation theory. Kelly and Glymour are skeptical of Bayesian confirmation theory because they believe that it does not square with the basic aims of scientific justification, particularly the cardinal aim of finding true answers to one's problems. They also maintain that Bayesians are not concerned with the proper nature of scientific justification, which they take to be showing the reliability and efficiency of the methods and strategies that are used to arrive at the truth. Instead, Bayesians are concerned with the updating of scientists' beliefs without regard for the reliability of their updating methods. Kelly and Glymour (2004) worry that "Bayesian methods assign numbers to answers instead of producing answers outright" (p. 112). For example, scientific theories can be, and mostly are, produced straight out without the accompanying posterior probabilities. Instead, they stress the importance of identifying problems and solving them by whatever means are appropriate. They acknowledge that there will be cases where Bayesian methods can do serviceable work.

The difficulties of deciding just what to think about Bayesianism are captured well by the ambivalence of the Bayesian philosopher of science, John Earman (1992), who thinks that it currently stands as our best philosophy of science. He confesses to being an enthusiastic Bayesian on Mondays, Wednesdays, and Fridays. But on Tuesdays, Thursdays, and Saturdays, he holds doubts about the totalizing ambitions of Bayesianism and, indeed, whether it can serve as a proper basis for scientific inference. Faced with such difficulties, he suggests that it is probably prudent to settle for a contextual application of Bayesian thinking. For example, in particular domains such as medical diagnosis, where the relevant probabilistic information is often available, scientists sometimes appeal to the Bayesian corpus to justify the selective use of its methods. By contrast, in domains where the evaluation of explanatory hypotheses and theories are of primary concern, scientists have, for good reason, often employed something like inference to the best explanation. Like it or not, the intending Bayesian scientist will have to

consult the relevant philosophical literature, among other methodological literatures, to furnish an informed justification for their Bayesian practices.

A Neo-Popperian Philosophy of Bayesian Statistics

So far in this chapter, I have concentrated on that part of Bayesianism known as Bayesian confirmation theory. Although this philosophical theory makes little explicit contact with standard Bayesian statistical theory and practice, it nonetheless provides the core of the philosophy that is consistent with the traditional view of Bayesian statistics. In this section, I turn my attention to an alternative philosophy of Bayesian statistics, neo-Popperian philosophy, that fits well with a different set of Bayesian statistical practices.

An alternative philosophy of Bayesian statistics was recently formulated by Gelman and Shalizi (2013; see also Gelman, Meng, & Stern, 1996). Their work provides a systematic and principled philosophical justification for a distinctive approach to current Bayesian statistical modeling practices. As will be seen, it has been influenced by the prominent philosopher of science, Sir Karl Popper—hence the label *neo-Popperian*.

Gelman and Shalizi reject the received philosophy of Bayesian statistics because they believe it fails to square with a current approach to Bayesian statistical modeling practices. Indeed, they claim that the prevalence of that philosophy hinders the development of Bayesian modeling. The received philosophy to which they object has the following primary characteristics: Bayesian statistical inference is regarded as a formal inductive process, whereby one learns about the general by reasoning from the particular; the researcher provides a subjective prior probability estimate that the model under examination is true; the researcher then smoothly updates his or her belief in the truth of the model through iterative use of Bayes's theorem to arrive at a posterior probability estimate of the model's truth; the model is then evaluated in relation to competing models, solely in terms of its posterior probability, which represents the Bayesian agent's degree of belief in the truth of the model. The so-called Bayes factor described previously is one popular method for undertaking comparative model evaluation.

Gelman and Shalizi's alternative philosophy is significantly shaped by Popper's (1969) view that scientific propositions are to be submitted to repeated criticism in the form of strong empirical tests. For them, best Bayesian statistical practice involves formulating models using Bayesian statistical methods and then checking them through attempts to falsify and modify those models.

In rejecting the orthodox Bayesian view that statistical inference is inductive inference, Gelman and Shalizi maintain that Bayesian inference is deductive inference. For them, the important process of model-based data analysis involves model checking, which they see as a deductive process that runs from model assumptions to conclusions about models. In this, they follow Popper, who maintains that there is no such thing as inductive inference. However, unlike Popper, Gelman and Shalizi allow for inductive inference, nested within deductive inference.

In a further contrast with Bayesian orthodoxy, Gelman and Shalizi maintain that prior probabilities are not personal beliefs held by agents; instead, they are assumptions that are postulated as a part of the hypothesized model, and they can be rejected or altered, if necessary, as part of the model's revision. In addition, models are not considered as true, or even approximately true. They are deemed false, as are all models. Instead, models are regarded as sets of assumptions that can be falsified, or modified, if their predictions do not square with the relevant data.

The final feature of Gelman and Shalizi's philosophy that I mention here is that it breaks with the standard Bayesian practice of deriving a probability estimate as the sole gauge of a model's worth. Instead, one attempts to falsify the model by carrying out what are termed "posterior predictive checks." These involve comparing simulated data based on the correctness of the model with the obtained empirical evidence in a goodness-of-fit exercise. In this way, one compares the model of interest with the data, without regard for other candidate models. The comparison is often done visually rather than by tests of statistical significance.

Popper and the Hypothetico-Deductive Method

Gelman and Shalizi maintain that formulating, checking, and revising models is best understood as a sophisticated form of hypothetico-deductive inference. For them, sophisticated

hypothetico-deductivism involves the adoption of a Popperian falsificationist view of the hypothetico-deductive method, with its emphasis on strong tests. Modeling, on this view, is a sequence of conjectures, refutations, and new, or modified, conjectures. It should be noted that Gelman and Shalizi are selective in their use of Popper. For example, they do not adopt his overarching theory of critical rationalism, his falsifiability criterion for demarcating science from nonscience, or his confirmation-theoretic notion of corroboration. However, consistent with Popper's view that there is no logic to scientific discovery, Gelman and Shalizi do not offer a methodological account of how models are formed, only how they are tested. Consistent with Popper's conception of hypothetico-deductive inquiry, they limit themselves to following Popper's injunction that one should engage in repeated strong testing of hypotheses about models, along with a commitment to the view that this should be done, principally by exploiting deductive inference.

In choosing to adopt a Popperian view of scientific method, Gelman and Shalizi explicitly reject the account of confirmation promoted by Carl Hempel (1965). Hempel proposes the idea that, in scientific confirmation, hypotheses are confirmed by discovering their positive instances. In his formalization of this idea, Hempel requires that the evidence entails the development of the relevant hypothesis with respect to the domain of the evidence. This contrasts with hypothetico-deductive inference, in which the evidence is deductively entailed by the hypothesis, and confirmation occurs through successful predictive testing.

It is important to point out that there are a number of sophisticated variants of hypothetico-deductive method available that overcome the limitations of the earlier simplified accounts. Gelman and Shalizi's version can be considered one of them. Sprenger (2011) provides a useful overview, and defense, of modern thinking about the hypothetico-deductive method.

Further, Sprenger (2013) proposes an account of confirmation that unifies Hempel's insight that hypotheses are confirmed by their instances and the core hypothetico-deductive idea that hypotheses are confirmed by their successful predictions. This modern hybrid account of hypothetico-deductive confirmation has an important advantage over that outlined by Gelman and Shalizi: It allows for an objective notion of inductive support,

which I believe that Gelman and Shalizi's model-testing strategy requires. At the same time, it features strong hypothetico-deductive testing of a falsificationist kind, and it allows for the piecemeal testing of entire theories, rather than their wholesale rejection. Both of these are desirable features of scientific modeling for Gelman and Shalizi.

In addition to drawing from Popper, Gelman and Shalizi make brief heuristic use of some of Thomas Kuhn's (1970) ideas about science. They suggest that Kuhn's distinction between normal and revolutionary science is somewhat analogous to their distinction between learning within a Bayesian model and checking the model either to discard or to expand it. However, they caution about pushing the analogy too far, correctly pointing out that most model checking and reformulation is puzzle-solving work, not revolutionary change, that takes place within a single paradigm. I think this disanalogy renders a serious appeal to Kuhn's theory of science as largely inappropriate for their particular philosophy, as I believe it is for the social sciences more generally.

Modes of Scientific Inference

As already noted, Gelman and Shalizi follow Popper in declaring that deductive inference is all there is to scientific inference. For them, this allows for the strong testing of Bayesian models by constantly checking them via their deductively derived predictions. However, importantly, and unlike Popper, they maintain that informative accounts of inductive reasoning, as they occur in science, will be material rather than formal. That is to say, they will be local rather than global in nature, in that the premises and conclusion of inductive arguments will contain reference to context-specific, contingent matters of fact (Norton, 2003). Furthermore, Gelman and Shalizi acknowledge that inductive statistical inferences to unobserved cases can be drawn on a background of deductive models. So, it would seem that, for them, science admits both deductive and inductive modes of reasoning.

I would go further than Gelman and Shalizi and claim that, in addition to deductive and inductive inference, science makes heavy use of a third type of inference known as abductive inference. Moreover, I believe that this form of inference could serve an important methodological role in Gelman and Shalizi's view

of model generation and model revision. Briefly, abductive inference is explanatory inference, and in science it involves reasoning about hypotheses, models, and theories in a manner that explains the relevant facts (e.g., Haig, 2014; Magnani, 2001). There are different species of abductive reasoning, having to do with the generation, modification, and appraisal of hypotheses and theories. For example, as will be seen in Chapter 6, the statistical method of exploratory factor analysis involves the abductive generation of latent factors to explain patterns in multivariate data. Further, inference to the best explanation, briefly mentioned by Gelman and Shalizi, is an abductive approach to theory appraisal, in which explanatory reasoning forms the basis for evaluating rival theories.

Although Gelman and Shalizi describe their modeling philosophy as falsificationist in nature, it goes beyond the strictures of Popper's view of the matter. When a model, or a component of a model, is confronted with negative evidence, the model, or its relevant parts, can be revised by means other than straight rejection or elimination. Often, this modification of a hypothesis will be seen to plausibly explain the anomalous data. For this reason, when scientists engage in such model revision, they employ abductive reasoning, whether they know it or not. Thus, it would seem that Gelman and Shalizi's account of Bayesian modeling requires an extension to include abductive inference to account for standard practices of model revision.

One final comment on the inference forms involved in modeling is in order. Gelman and Shalizi regard the process of checking and ruling out possible misspecifications of a model as consistent with the strategy of eliminative induction. However, in this context, they think the word *induction* is a misnomer, and they enlist the support of Kitcher (1993) in maintaining that the strategy really embodies a deductive argument. However, it should be noted that Kitcher is concerned with the successive elimination of actual theories that rival the theory of interest, not with successive checks for possible inconsistencies in a single model. I think it is clear that both inductive and deductive eliminative strategies are used in science, and that because of the uncertainties in social science research, the aspect of model checking referred to here by Gelman and Shalizi is more realistically construed as an inductive strategy.

The Value of Gelman and Shalizi's Philosophy

Gelman and Shalizi have provided the statistical fraternity with a philosophically informed perspective on Bayesian modeling that does justice to the building and revision of models through strong tests. Their Popper-inspired emphasis on strong tests of models is a welcome antidote to the reluctance of traditional Bayesian researchers to take model checking seriously.

Although predominantly Popperian in flavor, Gelman and Shalizi suggest that their philosophy can be strengthened by incorporating insights from Kuhn and Lakatos and the suggestive work of more recent philosophers. Their neo-Popperian philosophy, then, should be understood as a philosophy-in-the-making.

A heartening attitude that comes through in Gelman and Shalizi's expression of their philosophy is the firm belief that a philosophy of statistics is an important part of statistical thinking. Their work makes clear that philosophy can have a direct impact on statistical practice. Given that statisticians operate with an implicit philosophy, whether they know it or not, it is better to avail oneself of an explicitly thought-out philosophy that serves practice in useful ways. In this regard, they are at one with Mayo and Spanos's error-statistical philosophy, which is claimed in Chapter 3 to offer a strong philosophical aid to the understanding of tests of statistical significance and other frequentist statistical methods.

Gelman and Shalizi have done the statistical community considerable service by showing how Bayesian data analysis and modeling practices can be underwritten by different philosophies of science. In their work, these include inductive and deductive philosophies. Moreover, it is important to emphasize that the actual approach chosen by Gelman and Shalizi is but one of several that might be taken. For example, I have briefly suggested that a philosophy of abductive inference can further advance our understanding of model building. Plausibly, then, the philosophy of data analysis cannot be restricted solely to one of the major theories of scientific inference. The process of data analysis could be argued to feature elements that are reminiscent of several theories in the philosophy of science, including inductive, deductive, and abductive accounts. However, data analysis is also guided and constrained by strongly pragmatic concerns, ranging from the available money and time to the computational resources of computers, which are alien to

Bayesianism, as well as to the type of hypothetico-deductive reasoning strategies that Gelman and Shalizi enlist.

Conclusion

As this selective discussion of Bayesian confirmation theory and statistics makes clear, Bayesianism is a highly contested perspective, and what I say in this chapter on the topic should, therefore, be taken primarily as a stimulus package for readers to develop their own views about it.

At a general level, Bayesianism comprises two major overlapping strands: confirmation theory and statistics. However, these two strands are mostly presented independently of each other. Further progress in the foundations of Bayesian confirmation theory is likely to be made if methodologists consider it in the light of statistics. For its part, Bayesian statistics stands to be enriched by a more concerted effort to incorporate insights from Bayesian confirmation theory into its fold.

Bayesian thinkers differ widely among themselves regarding how Bayesianism should be characterized and evaluated. For example, there is a major contrast between "subjectivist" Bayesians and "objectivist" Bayesians; there are many different proposals for dealing with the problem of how to make good prior probability estimates; and there are those who think that Bayesian and frequentist statistical practices can be fruitfully combined. The considerable variety of Bayesian alternatives on offer needs to be more widely appreciated.

Deciding just how much sound scientific practice is, in fact, captured by the Bayesian apparatus is a major challenge. For example, although scientists often talk about accepting theories, and the strength of evidence for and against them, they seldom talk about their probabilities. Additionally, some critics (e.g., Glymour, 1980) think that Bayesian theory is too far removed from the history of scientific practice to be genuinely informative, though it should be acknowledged that, in recent years, some Bayesian analyses of episodes in the history of science have been carried out. There is also the major challenge of understanding the normative force for science of Bayesian thinking: Should we be Bayesians, and if so, when and how?

Recommendations to adopt Bayesianism across the board should be resisted. The folly of the near-universal adoption of traditional null hypothesis tests of statistical significance should not

be repeated by replacing them in a totalizing manner with Bayesian statistical alternatives. It seems prima facie implausible that the simple piece of mathematics that is Bayes's theorem could provide the basis for deep insights into all of the rich variety of inferential practices that are to be found in successful science. Bayesian methods are likely to best serve science in local, domain-specific ways, much as other methods do.

In deciding whether to adopt a Bayesian position on statistical inference, it should be kept in mind that one does not have to embrace a general Bayesian theory of scientific confirmation rather than, say, a modern hypothetico-deductive alternative. One might be a Bayesian when dealing with problems of statistical inference, but remain wedded to a defensible hypothetico-deductive conception of scientific method. Or, more plausibly, one might employ Bayesian statistical methods when concerned with inferential problems about hypotheses for which we have the relevant probabilistic information, but otherwise adopt a nonprobabilistic count of theory evaluation such as Thagard's (1992) theory of explanatory coherence. The general point to be made here is that Bayes's theorem can help us deal with some problems of scientific inference, but, clearly, a great deal of scientific work will be done with the use of other methods, some of them statistical and some of them not.

Further Reading

Peter Godfrey-Smith's philosophy of science textbook (*Theory and Reality: An Introduction to the Philosophy of Science.* Chicago, IL: University of Chicago Press, 2003) contains a short, accessible overview of Bayesian confirmation theory and a promissory note on an alternative outlook.

Robert Nola and Howard Sankey's *Theories of Scientific Method: An Introduction* (Montreal, Canada: McGill-Queen's University Press, 2007) has a more advanced and extended treatment of Bayesianism that is sympathetic to that outlook.

Colin Howson and Peter Urbach's *Scientific Reasoning: The Bayesian Approach* (3rd ed. La Salle, IL: Open Court, 2006) provides a lucid and thorough-going Bayesian analysis of scientific inference by authors who are strongly committed to the approach.

John Earman's *Bayes or Bust*? *A Critical Examination of Bayesian Confirmation Theory* (Cambridge, MA: MIT Press, 1992), true to its subtitle, offers a sympathetic, but critical, examination of Bayesian confirmation theory. Although he thinks Bayesianism is currently the best approach to the philosophy of science, and holds the most promise for an adequate general theory of confirmation, he delivers a mixed verdict regarding its successes and failures to date.

In an extended critical review of Earman's book, Malcolm Forster argues, against Earman, that Bayesianism is not the most promising comprehensive theory of scientific confirmation. See his, "Bayes and Bust: Simplicity as a Problem for a Probabilistic Approach to Confirmation" (*British Journal for the Philosophy of Science, 46*, 399–424, 1995).

Clark Glymour offers an incisive critique of the subjective Bayesian theory of confirmation in his book, *Theory and Evidence* (Princeton, NJ: Princeton University Press, 1980).

In a suggestive contribution to Bayesian methodology, Wesley Salmon (1990) argues that Thomas Kuhn's five criteria for a good theory (accuracy, consistency, scope, simplicity, and fruitfulness) can be incorporated within a Bayesian approach. See his essay, "Rationality and Objectivity in Science, *or* Tom Kuhn Meets Tom Bayes" (in C. W. Savage, Ed., *Scientific theories*, pp. 175–204. Minneapolis, MN: University of Minnesota Press, 1990).

A classic paper that brought the Bayesian outlook in statistics to psychology is Ward Edwards, Harold Lindman, and Leonard Savage's "Bayesian Statistical Inference for Psychological Research" (*Psychological Review, 70*, 193–241, 1963).

Two accessible books on Bayesian data analysis written for social and behavioral scientists are David Kaplan's *Bayesian Statistics for the Social Sciences* (New York, NY: Guilford, 2014) and John Kruschke's *Doing Bayesian Data Analysis: A Tutorial With R and Bugs* (Boston, MA: Academic Press, 2011).

A more advanced textbook on Bayesian data analysis is Andrew Gelman at al.'s well-known *Bayesian Data Analysis* (2nd ed. London, England: Chapman and Hall, 2003).

Gelman is the author, with Cosma Shalizi, of the important article, "Philosophy and the Practice of Bayesian Statistics" (*British Journal of Mathematical and Statistical Psychology, 66*, 8–38, 2012) discussed in this chapter.

Zoltan Dienes compares orthodox and Bayesian approaches to statistical inference in his article, "Bayesian Versus Orthodox Statistics: Which Side Are You On?" (*Perspectives on Psychological Science, 6*, 274–298, 2011). He expresses a clear preference for the Bayesian approach and shows how to implement Bayesian hypothesis testing in practice, with an emphasis on calculating Bayes factors.

An important review of Bayes factors as an approach to Bayesian hypothesis testing is presented in Robert Kass and Adrian Raftery's "Bayes Factors" (*Journal of the American Statistical Association, 90*, 773–795, 1995).

An instructive article, with a specific focus on the philosophy of Bayes factors as a means of quantifying statistical evidence, is Richard Morey, Jan-Willem Romeijn, and Jeffrey Rouder, "The Philosophy of Bayes Factors and the Quantification of Statistical Evidence" (*Journal of Mathematical Psychology*, 72, 6–18, 2016).

References

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*, 521–543.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society (London)*, *53*, 370–418.

Bunge, M. (2008). Bayesianism: Science or pseudoscience? *International Review of Victimology*, *15*, 165–178.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory.* Cambridge, MA: MIT Press.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–241.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–760.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8–38.

Glymour, C. (1980). *Theory and evidence.* Princeton, NJ: Princeton University Press.

Good, I. J. (1971). 46656 varieties of Bayesians. *American Statistician*, *25*, 62–63.

Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *American Journal of Psychology*, *122*, 219–234.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences.* Cambridge, MA: MIT Press.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York, NY: Free Press.

Horwich, P. (1982). *Probability and evidence.* Cambridge, England: Cambridge University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). La Salle, IL: Open Court.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Clarendon Press.

Kelly, K. T., & Glymour, C. (2004). Why probability does not capture the logic of scientific justification. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 94–114). Malden, MA: Blackwell.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. New York, NY: Oxford University Press.

Kruscke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam, the Netherlands: Elsevier.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Kyburg, H. (1992). The scope of Bayesian reasoning. In *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association* (Vol. 2, pp. 139–152). Chicago, IL: University of Chicago Press.

Laudan, L. (1981). *Science and hypothesis.* Dordrecht, the Netherlands: Reidel.

Leplin, J. (1997). *A novel defense of scientific realism.* Oxford, England: Oxford University Press.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London, England: Routledge.

Magnani, L. (2001). *Abduction, reason, and science: Processes of discovery and explanation*. New York, NY: Kluwer/Plenum.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science*, *54*, 553–567.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*, 647–670.

Popper, K. R. (1969). *Conjectures and refutations: The growth of scientific knowledge* (3rd ed.). London, England: Routledge & Kegan Paul.

Psillos, S. (2004). Inference to the best explanation and Bayesianism. In F. Stadler (Ed.), *Induction and deduction in the sciences* (pp. 83–91). Dordrecht, the Netherlands: Kluwer.

Rosenkrantz, R. D. (1977). *Inference, method, and decision: Towards a Bayesian philosophy of science.* Dordrecht, the Netherlands: Reidel.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–391). Hillsdale, NJ: Erlbaum.

Salmon, W. C. (1966). *The foundations of scientific inference*. Pittsburgh, PA: University of Pittsburgh Press.

Salmon, W. C. (1970). Bayes's theorem and the history of science. In R. H. Stuewer (Ed.), *Historical and philosophical perspectives of science: Minnesota studies in the philosophy of science* (Vol. 5, pp. 68–86). Minneapolis, MN: University of Minnesota Press.

Salmon, W. C. (1990). Rationality and objectivity in science, *or* Tom Bayes meets Tom Kuhn. In C. W. Savage (Ed.), *Scientific theories: Minnesota studies in the philosophy of science* (Vol. 14, pp. 175–204). Minneapolis, MN: University of Minnesota Press.

Savage, L. J (1964). The foundations of statistics reconsidered. In H. Kyburg & H. Smokler (Eds.), *Studies in subjective probability* (pp. 173–188). New York, NY: Wiley.

Shimony, A. (1993). *Search for a naturalistic world view: Volume 1: Scientific method and epistemology*. Cambridge, England: Cambridge University Press.

Sprenger, J. (2011). Hypothetico-deductive confirmation. *Philosophy Compass*, *6*, 497–508.

Sprenger, J. (2013). A synthesis of Hempelian and hypothetico-deductive confirmation. *Erkenntnis*, *78*, 727–738.

Stigler, S. M. (1983). Who discovered Bayes's theorem? *American Statistician*, *37*, 290–296.

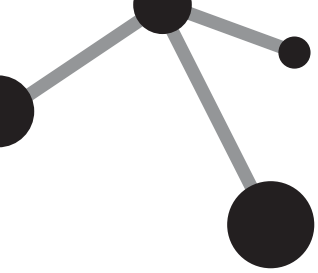
Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ: Princeton University Press.

Thagard, P. (2000). *Coherence in thought and action.* Cambridge, MA: MIT Press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Weisberg, J. (2009). Locating IBE in the Bayesian framework. *Synthese*, *167*, 125–143.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.



5

META-ANALYSIS

I think that meta-analysis is one of the most important methodological contributions of this generation of psychologists, arguably the most important.

—P. E. Meehl, 1990, p. 242

In the majority of cases where [meta-analysis] is being used, . . . it muddies the waters, disregards the problems, and leads to meaningless conclusions that are likely to hamper proper scientific research.

—H. J. Eysenck, 1984, p. 58

Introduction

An important part of the evaluation of the state of knowledge in science involves reviews of the literature in its many fields. In this regard, it is noteworthy that in the last 40 years there has been an enormous increase in attention given to the nature and place of literature reviews in science, along with concerted efforts to conduct them in a more rigorous and systematic fashion.

Until the 1980s, literature reviews in science were mostly narrative in form, drawing conclusions about multiple studies on a topic

in a qualitative rather than quantitative manner. Although narrative literature reviews occupy an important place in the reviewing process within science, they have been frequently criticized as casual, severely selective, and unable to portray cumulative knowledge (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Glass, 1976; Light & Smith, 1971). However, the simple quantitative option of vote taking from box score tallies of statistical significance test outcomes has been faulted for its failure to acknowledge the methodological asymmetry between confirmation and refutation (refutation being more decisive than confirmation) and for its bias in favor of large-sample studies for which the significant outcomes are largely a function of statistical power (e.g., Meehl, 1978).

In the face of burgeoning and fragmented research literatures displaying conflicting results, meta-analysis has developed as a systematic and objective alternative to the customary integration methods of narrative literature reviews and vote counting of significance test outcomes. The development and widespread use of meta-analytic procedures to integrate or synthesize the results of empirical studies in many areas within the behavioral, medical, and social sciences stands as one of the most striking methodological developments of the last four decades. Its rapid uptake has been described graphically as "the meta-analytic big bang" (Shadish & Lecy, 2014).

Meta-analysis is an approach to data analysis that involves the quantitative, or statistical, analysis of data analyses from a number of existing primary studies in a common domain. At its simplest, meta-analysis involves computing the average effect size for a group of studies. For the originator of modern meta-analysis, Gene Glass, the effect size measure is the standard score obtained by subtracting the mean of the control group from that of the treatment group and dividing this difference by the standard deviation of the control group (Glass, 1976). This is done for each of the relevant dependent variables in each study. The effect sizes are then summed and divided by the total number of effects to obtain the average effect size. Of course, there is much more to Glassian meta-analysis than this (Glass, McGaw, & Smith, 1981), but my focus in this chapter is on the underlying rationale for meta-analysis.

Meta-analysis comes in a variety of forms (e.g., Bangert-Drowns, 1986; Borensteinet al., 2009). Prominent among them are Glassian meta-analysis and its technical advancement by Hedges

and Olkin (1985); Schmidt and Hunter's (2015) psychometric approach to meta-analysis; and Rosenthal's (1978) combined probability method. These, and other approaches to meta-analysis, are discernibly different from one another, but their differences are not relevant to my treatment in this chapter.

Despite the claimed advantages of meta-analysis, and its current popularity as an approach to literature reviews, a number of different types of criticism have been leveled against it. For example, Slavin (1984) has argued that the use of meta-analytic procedures in the field of education constitutes a retrograde step in the art of research integration. Similarly, Ioannides (2016) recently showed that the plethora of meta-analyses in biomedical science is often "redundant, misleading, and conflicted" (p. 485). Others (e.g., Bruno & Ellett, 1988; Cook & Leviton, 1980; Erwin, 1984; Sohn, 1996) have pointed out what they take to be methodological limitations of the approach, while at a meta-theoretical level, the arch enemy of meta-analysis, Hans Eysenck (1984), has argued that the meta-analytic enterprise is unscientific, and constitutes "an abuse of research integration" (p. 41).

My primary concern in this chapter is with the conceptual foundations of meta-analysis. My examination is selective and centers on large-scale issues having to do with meta-analysis and the nature of science. I give considerable space to presenting the conception of inquiry embodied in the underlying rationale of Glass's approach to meta-analysis. I will be concerned as much with making his rationale known, as I will be with evaluating it. I then examine David Sohn's provocative argument that meta-analysis is not a legitimate vehicle of scientific discovery. After that, I consider the role of meta-analysis in relation to the processes of phenomena detection and scientific explanation. In doing so, I examine the extent to which meta-analysis can properly be said to contribute to scientific progress.

Glass's Rationale for Meta-Analysis

Somewhat surprisingly, evaluations of Glass's approach to meta-analysis have shown little regard for his underlying rationale (Glass, 1972; Glass & Kliegl, 1983). Glass rightly claims that many misunderstand his meta-analyses of outcome research because they fail to be cognizant of the rationale he provides. This failure

is offered by him as the main reason for the widespread misunderstanding of Smith, Glass, and Miller's (1980) original meta-analysis of psychotherapy outcome studies. Unfortunately, inattention to the rationale has deprived methodologists and researchers of the opportunity to reflect on an important contribution to the conceptual foundations of meta-analytic methodology—an omission that this chapter begins to correct.

Scientific and Evaluative Inquiry

The core of Glass's rationale for meta-analysis involves drawing a distinction between scientific, or elucidatory, and evaluative inquiry (Glass, 1972; see also Smith et al., 1980). Glass's position is that researchers as scientists are concerned to satisfy their curiosity by seeking truthful conclusions in the form of theories comprising explanatory laws. By contrast, evaluators undertake research on behalf of a client, which is aimed at producing useful decisions based on descriptive determinations of the worth of particular products or programs. Importantly, for Glass, the meta-analysis of outcome studies properly involves the integration of the empirical results of evaluative research only.

Glass differentiates scientific from evaluative inquiry in respect to a number of basic contrasts. I present four of the most important of them and offer some evaluative comments about their tenability. Generally speaking, I favor the view that evaluative inquiry can be seen as a form of scientific inquiry, not something fundamentally different from it.

Motivation of the Inquirer. According to Glass, scientific inquiry is undertaken largely to satisfy the curiosity of the researcher and, to this end, involves the construction of theories. By contrast, the researcher's basic concern in conducting evaluative inquiry is to help solve a client's practical problem. The findings of meta-analytic research constitute an important source of evidence for helping solve such problems, but they are not to be used as the empirical foundation for building explanatory theories.

Comment. It is doubtful whether the satisfaction of a researcher's curiosity, the construction of theories, or any other aim (e.g., truth, understanding, control) should be taken as the primary concern of

science. This is because science, carried out by human agents and embedded in social institutions, attends to multiple goals that are pursued simultaneously. Moreover, the aims of science are genuinely problematic and are provisionally arrived at by debate within science's critical community (Hooker, 1987). Important among these aims are epistemic goals that include, for example, the detection of empirical phenomena (often in the form of robust empirical generalizations) and the construction of theories to coherently explain those phenomena. Further, science will legitimately seek nonepistemic goals, such as pragmatic utility and risk assessment, when engaged in policy formulation and the application of scientific knowledge. Later, I suggest that, in contrast to Glass, meta-analysis can be properly viewed as an important means by which we can discover empirical phenomena in science.

Laws and the Particular. Glass (1972) briefly invokes the popular distinction between nomothetic and idiographic research to differentiate further scientific and evaluative inquiry. For him, scientific inquiry involves the search for laws understood as statements of general relationship among variables or phenomena, whereas evaluation involves the description of the value, or values, of a particular thing.

Comment. This contrast between nomothetic and idiographic forms of inquiry is clearly based on the widely held view that causal laws are universal, or widely applicable, empirical regularities. However, the nature of laws is a contested matter, and it is just as defensible to think of causal laws as the causally necessary activity of generative mechanisms rather than their conditions of activation or expressions of effect (Bhaskar, 1978; Harré & Madden, 1975). On this view, it is a contingent matter whether or not the mechanisms happen to be in a closed system, like an experiment, in which they can produce empirical regularities. A law does not cease to exist in an open system just because its empirical manifestations are absent. It is just that manifestations are typically altered or checked by the work of other causal mechanisms in an open system.

By taking causal laws to involve the natural necessity of causal mechanisms rather than the scope of empirical regularities, we can question Glass's particular use of the popular distinction

between nomothetic and idiographic inquiry for wrongly supposing that causal laws and claims about the particular have to be to be considered as alternatives. Because laws can be construed as a matter of causal necessity, nomothetic inquiry can be either idiographic or universal in nature. A science of the particular is a perfectly proper project, as for example with the study of individual lives using autobiographical methods. Indeed, there is a good argument that, in psychology, idiographic research should receive as much weight as so-called individual differences research (Molenaar, 2004). Moreover, to endorse a study of the particular is not to foreclose the possibility that the future comparative study of individual lives may well reveal deep-structural generalizations or perhaps even universals (cf. DeWaele & Harré, 1979). Nomothetic and idiographic inquiry are complementary rather than mutually exclusive.

The Role of Explanation. According to Glass (1972), science involves the continual search for subsurface explanations of empirical phenomena. Evaluative inquiry, on the other hand, does not seek explanations: "A fully proper and useful evaluation can be conducted without producing an explanation of why the product or program being evaluated is good or bad or how it operates to produce its effects. . . . [It] is usually enough for the evaluator to know that something attendant upon the [product or program] is responsible for the valued outcomes" (pp. 5–6). Glass's position seems to be that, even though program treatments can be causally responsible for their measured outcomes, it matters little that knowledge of this gleaned from evaluation studies does not tell us how programs produce their effects because such knowledge is not needed for policy action.

Comment. Glass is undoubtedly correct in asserting that scientists are centrally concerned with the construction of causal theories to explain empirical phenomena for this is the normal way in which they achieve understanding of the empirical regularities they discover (Haig, 2014). However, he is wrong to insist that proper evaluations can, or should, deliberately ignore knowledge of underlying causal mechanisms. The reason for this is that the effective implementation and alteration of social programs will often benefit from knowledge of the relevant causal mechanisms

involved (Gottfredson, 1984), and strategic intervention with these in mind is sometimes the most effective way to bring about social change. Even though the relevant causal mechanisms will typically be unobserved, appeal to knowledge of such mechanisms will nevertheless increase our understanding of relevant matters and help us implement change.

Truth and Social Utility. This is probably the major contrast for Glass. He asserts that scientific inquiry characteristically attempts to assess the truth of knowledge claims, whereas evaluative inquiry attempts to gauge the worth of things. Glass takes truth to comprise the empirical validation and logical consistency of knowledge claims, while worth is understood as social utility. He acknowledges that truth is highly valued and worthwhile, but, this point aside, he insists that the contrast between truth and utility effectively helps to distinguish science from evaluation.

Comment. It is important to appreciate here that Glass fails to make an epistemic distinction that is crucial to a satisfactory understanding of the nature of science: In identifying truth with empirical adequacy and logical coherence, Glass has conflated the epistemic notions of truth and justification. Truth is best understood as correspondence with reality, where it functions as a guiding ideal for science. As such it is a highly valued, though unattained, goal that helps us make sense of science as an attempt to represent and intervene in the world. However, truth is only accessible indirectly by way of the various criteria we use to justify and accept theories. Empirical adequacy and logical coherence are in fact two such criteria. They do not constitute truth itself but instead function as surrogates for truth (Haig & Borsboom, 2012; Hooker, 1987).

The sketch I have presented in response to the tenability of Glass's four contrasts takes science to be an aim-oriented human endeavor that seeks to construct truthful causal explanatory theories of both the particular and the general. It is a view of science that rejects Glass's strong distinction between scientific and evaluative research.

I turn now to consider Glass's views about the nature of methodology, which underwrites his conception of meta-analysis.

The Nature of Methodology

Methodology Is Empirical

An important part of the rationale for Glassian meta-analysis involves adopting a conception of methodology as a substantive empirical undertaking. According to Glass, critics have often misunderstood his conception of meta-analysis because they have failed to appreciate that it embodies a methodology of this sort. He claims that, for any given empirical domain, a methodology combines with an object field, and a taxonomy, to give that domain its basic structure (Glass & Kliegl, 1983). For Glass, none of these three components is given to us a priori as a product of logic. Instead, they are chosen for both arbitrary and historical reasons. That is to say, methodologies are selected and developed partly as a response to the structure and pragmatic needs of society. For example, Fisherian (agrarian) experiments are said to embody principles that grow out of the demand for control made by a technological society. Glass follows Meehl (1978) in claiming that methodological assumptions are genuinely refutable conjectures. Indeed, it should be emphasized that one of the main functions of Glassian meta-analysis is to undertake the empirical investigation of such assumptions as part of its own object field. Glass believes that the most serious criticisms of meta-analysis lose their force when they are examined from the standpoint of empirical methodology.

I believe that Glass is right to criticize the influential conception of methodology for its a priori status, but that he is wrong to suggest that methodology is solely an empirical enterprise. Viewing methodology as a priori knowledge is dubious because the notion of a priori knowledge is itself highly questionable. The a priori categories of analytic truth, synthetic a priori truth, logical truth, and mathematical truth have all been subjected to serious criticism within philosophy (e.g., Haack, 1974; Kitcher, 1983; Quine, 1953). But, although one can accept Glass's claim that methodological statements are genuinely refutable conjectures, it does not follow that methodological assertions are evaluated solely on empirical grounds. The reason for this is that, in science, much procedural knowledge, no less than substantive knowledge, has the status of warranted conjectural theory and that, broadly speaking, both kinds of knowledge are validated using the methods of science

(Haig, 2014; Hooker, 1987). Because substantive scientific theories are underdetermined by the relevant data, they are additionally evaluated on superempirical dimensions such as explanatory power, systemic worth, and fruitfulness. We should not expect it to be any different with our methodological theories. To be sure, empirical evidence will have an important bearing on assessing the soundness of methodological claims, but these assessments will be inconclusive without invoking appropriate superempirical criteria. Additionally, I note that, where Glass admonishes researchers for engaging in what he thinks are a priori methodological debates, it may very well be the case that some of the disputes are really a posteriori intertheory debates about contingent matters of fact.

Indeed, Glass and his associates (Glass et al., 1981; Smith et al., 1980) have repeatedly emphasized that meta-analysis recommends itself over traditional review procedures because of its objectivity. This is said to be achieved by adopting judgment strategies in meta-analysis that will prevent biases from entering into the results it produces.

Meta-Analysis and Policy

Glass and Kliegl (1983) maintain that it is naïve to believe that rational policy decisions must be based on relevant knowledge from well-established theories. They echo Meehl's (1978) judgment that "most so-called 'theories' in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless" (p. 806). By reinterpreting such theories as the modest products of evaluative research, and submitting them to meta-analysis where appropriate, Glass and Kliegl (1983) believe that useful knowledge can be provided for decision-makers. In this way, they believe they can overcome researchers' habitual tendency to engage in "partisan squabbles and theoretical hot-dogging when attempting to inform policy makers" (p. 35).

However, meta-analysis has often failed in its attempt to establish clear judgments of pragmatic worth for policy makers. Different meta-analyses in the same subject area have often produced different results. For example, the constructive replication of the initial Smith et al. (1980) meta-analysis of psychotherapy outcome studies by Prioleau, Murdock, and Brody (1983) produced discrepant

conclusions. Because meta-analyses are unavoidably replete with human judgments over which researchers will differ, it is only to be expected that they will be unable to provide clients with unambiguous messages.

Also concerning policy, it is worth noting that Glass and Kliegl (1983) make inappropriate use of Habermas. They claim that "Habermas (1971) argued convincingly that the knowledge-constitutive interests that determine, in part, the selection of a certain methodology for science can be derived from the structure and pragmatic needs of the society in which the science exists" (p. 35). However, Habermas's (1971) critical-theoretic analysis of cognitive interests relies uncritically on an inappropriate empiricist theory of science (a point Habermas himself now concedes). Relatedly, Habermas's insistence that our knowledge-constitutive interests are somehow transcendental and a priori is implausible and clearly should be anathema to Glass and Kliegl.

This point aside, it is important to stress that methodologies and social institutions do relate to each other in mutually supporting ways (Unger, 1975). It can plausibly be argued that, to the extent that meta-analysis adopts a descriptive, atheoretical conception of inquiry, it helps to serve as a prop for our extant social institutions by providing them with conceptual resources that help maintain, rather than challenge, the status quo. One way in which meta-analysis reinforces the status quo stems from its encouragement of, and reliance on, narrowly focused primary studies.

Meta-analysis further reinforces the status quo by restricting its attention to outcome studies that focus on phenomenal appearances and refrain from considering underlying causes. This willingness to stop short of attempting to tell decent causal stories contributes to a general inability to regard educational programs and social institutions more generally as structurally problematic and results in an absence of coherent knowledge of the relevant causes, which would be the objects of strategic social change.

A third way in which meta-analysis reinforces the status quo stems from the fact that it is not critically aim oriented. By willingly accepting clients' goals, evaluation research employs meta-analysis as part of an instrumental rationality concerned to devise and follow efficient means to clients' ends. As such, meta-analytic methodology affords us neither the inclination nor the ability to challenge the goals of clients, programs, or social institutions.

Glass and Kliegl (1983) are undoubtedly correct in claiming that the sources of methodologies lie in the pragmatic interests of social groups and institutions, rather than in logic. However, they do not consider the extent to which methodologies and social conditions can mutually reinforce one another. They also do not appreciate the role of a critically aim-oriented conception of science in furthering our understanding of inquiry.

In the next three sections of the chapter, I consider additional aspects of the relationship between meta-analysis and the nature of science.

Meta-Analysis and Scientific Discovery

One of the originators of modern meta-analysis, Frank Schmidt (1992), has championed the view that meta-analysis is of central importance to the advancement of scientific knowledge. He entertains a revisionist model of possible (though not necessarily desirable) future science as

> a two-tiered research enterprise. One group of researchers will specialize in conducting individual studies. Another group will apply complex and sophisticated meta-analysis methods to those cumulative studies and will make the scientific discoveries. (p. 1180)

In a series of related articles, David Sohn (1995, 1996, 1997) strongly challenges the view that meta-analysis can serve as a proper means of scientific discovery in the manner suggested by Schmidt. There are two parts to his challenge: (a) He claims that the quality of empirical psychological studies that are used in meta-analysis is unacceptably low; and (b) he believes that meta-analysis is a form of scientific review, but it is not a genuine form of research. The two claims are related, in the sense that the poor quality of primary studies undermines the worth of meta-analyses (whether they be thought of as scientific research or scientific reviews). I concentrate on Sohn's argument that meta-analysis is not an important vehicle of scientific discovery. In doing so, I comment in passing on his reservations about the problems that he thinks beset mainstream psychological research.

Sohn (1996) questions the basic idea of meta-analysis as a standalone literature review capable of discovering truths, whereas

traditionally scientific discoveries were contained in the empirical findings of the primary studies themselves. For Sohn, the idea that meta-analytic literature reviews can make discoveries about nature rests on the assumption that the primary research literature is a proxy for nature. It is this assumption that he roundly rejects.

Noting the tendency of meta-analysts to paint a bleak picture of progress in twentieth-century psychology, Sohn (1996) suggests that although meta-analysis has been introduced to improve matters in this regard, it is in fact symptomatic of its poor progress. In his judgment, this lack of good progress is a consequence of psychology adopting a hypothesis-testing view of science. For Sohn, this view of science seeks knowledge by testing research hypotheses about the relationship of descriptive variables without regard for causal mediating variables. Essentially, the approach amounts to a hypothetico-deductive testing of outcome studies through use of methods such as tests of statistical significance and effect size measures. Sohn maintains that there are in fact two deleterious consequences of such an approach to research; one is the lack of agreement about outcomes, and the other is the absence of knowledge of the causal mechanisms that are responsible for those alleged outcomes. Although Sohn judges the second defect to be more serious, meta-analysis is indicted by Sohn for failing to remedy both defects.

However, Sohn supports his claim that meta-analysis does not produce demonstrable evidence for treatment effects in a curious way. He acknowledges that Smith et al.'s (1980) well-known meta-analytic treatment of the benefits of psychotherapy has been corroborated by subsequent meta-analyses, yet he maintains that this does not constitute evidence for replicable effects. He expresses a distrust of research that relies on statistical methods for making claims about replicable effects. This distrust appears to be founded in part on an extension of the view attributed to Lord Rutherford that if an experimental study requires statistics, then the experiment is in need of improvement. For Sohn (1996), "If one's science needs [meta-analysis], one should have done better science" (p. 243).

However, this idiosyncratic view of experimental inquiry flies in the face of widely accepted scientific practice. For example, detailed examination of both experimental and nonexperimental inquiry in science strongly supports the view that different parts of the various sciences from physics to psychology appropriately make

extensive use of statistical methods in the detection of empirical phenomena (Haig, 2009; Woodward, 1989). For their part, most statisticians today, influenced by John Tukey's pioneering work in the field, see their discipline as a science. Statistics would not exist as we currently know it unless it provided a necessary armament for science.

In this regard, it is worth noting that Sohn acknowledges the claim made by Hedges and Olkin (1985) that meta-analysis in some form or other has a long history of use in the hard sciences. Interestingly, Sohn states his disagreement with this position, but he does not argue against it. Had he done so, he might reasonably have been expected to counter Hedges's (1987) empirically based argument for the conclusion that exemplary practice in the soft science of psychology compares favorably with successful practice in the hard science of physics.

In Sohn's judgment, a more serious shortcoming of meta-analysis, and the hypothesis-testing research on which it is based, is that it fails to contribute to an understanding of the mechanisms that might afford an explanation of the empirical effects. However, to indict meta-analyses of the efficacy of psychotherapy, for example, for not contributing to a theoretical understanding of the therapeutic process, is to misconceive its basic purpose. Leaving to one side Glass's view that meta-analysis is an approach to evaluative research, I have suggested previously that the basic purpose of meta-analysis is to assist in the detection of empirical phenomena. This important part of scientific discovery is quite different in kind from the related kind of discovery that involves the construction of theories to explain the phenomena. Being different types of scientific discovery, they essentially make use of quite different research methods. Exploratory and confirmatory data analytic methods figure prominently in the process of phenomena detection. By contrast, abductive research methods are designed to help in the construction of explanatory theories (Haig, 2005, 2014). I have more to say about these two different types of discovery in the next section.

Sohn (1996) advances an extended argument that he believes discredits the idea that meta-analysis is a vehicle of scientific discovery. He presents the argument in summary form as follows:

> The literature is not a surrogate for nature, and literature study is not a substitute for nature study. Furthermore, the activity

> of the empirical scientist is of a fundamentally different kind from that of the literature reviewer. The published literature of psychology, because of publication bias, is a catchment for Type I errors, which are of two kinds, honest and dishonest. Efforts to obtain information about nature from this literature are likely to fail, if for no other reason, because of Type I errors. When a scientific proposition is true, this fact will be discernable, to the scientists working in the area, in the findings of empirical research. In such a case, a literature review will be redundant. Proposals for mitigating the effects of publication bias are difficult to assess and introduce complications not found in empirical science. None of these proposals has considered the problem of dishonest Type I errors, those due to researcher cheating. . . . In . . . the case of science [the process of seeking truth] is self-correcting. . . . The process of literature review, however, is not self-correcting. (p. 135)

Here Sohn challenges the assumption of meta-analysts that the research literature contains information about nature that can be mined by meta-analytic studies. This assumption, he maintains, takes the literature as a proxy for nature. However, Sohn reasons that because the primary literature in psychology contains a significant amount of Type I error resulting from the regular and improper use of tests of statistical significance, and is therefore not an accurate reading of nature, meta-analysis cannot take the results of primary studies at face value. In fact, the situation is worse than this, for many studies that do not reach an acceptable level of statistical significance are filed away in the drawers of researchers (the so-called file drawer problem; Rosenthal, 1979).

Sohn is well aware that meta-analytic methodology has resources designed to mitigate the effects of the publication bias just noted. He asserts that these are difficult to assess but he does not pursue the matter. However, without an attempt to consider the effectiveness of the methods for correcting publication bias, his argument on this point carries little weight. Instead, Sohn expresses the general worry that tests for publication bias introduce a factor of uncertainty that does not exist for good primary research. This it does, but publication bias is not the only source of research bias, and primary studies are subject to a range of different biases, such as investigator bias, sample bias, confirmation bias, and reviewer

bias (Schmidt & Hunter, 2015). And, these biases bring with them their own uncertainties.

Regarding the matter of uncertainty in science more generally, it should be said that science is constantly concerned with identifying, estimating, and controlling for ignorance and uncertainty and coping with the ensuing levels of doubt. This is inevitable because science is undertaken by fallible agents who work in imperfect institutions with limited, but useful, methods to help them study complex subject matters. Whatever their limitations, methods that correct for publication and other forms of bias are methods devised to help deal with these sources of uncertainty.

Because of space limitations, I have only traced the broad contours of Sohn's case against the worth of meta-analysis as a means of scientific discovery. However, I believe that Sohn's arguments against meta-analysis as a means of scientific discovery are not convincing.

Meta-Analysis and Phenomena Detection

As noted at the beginning of the chapter, meta-analysis is an approach to data analysis that involves quantitative analysis of the data analyses of primary empirical studies. By calculating effect sizes across primary studies in a common domain, meta-analysis helps us detect "ubiquitous positive effect[s]" (Schmidt, 1993, p. 1164) (or more accurately, general positive effects). As such, it is a prominent example of a distinctive use of statistical methods by behavioral scientists to engage in the process of phenomena detection. By using statistical methods to detect the existence of robust empirical regularities (the most common type of phenomena), meta-analysis can in fact be usefully viewed as a statistical approach to constructive replication. Constructive replication is undertaken to establish the extent to which findings hold across different methods, treatments, and occasions. It is a triangulation strategy employed to establish the generalizability of results identified by direct replication. Although meta-analysis is used quite widely in evaluation research and is thought by some to do explanatory work, it is in the descriptive-cum-generalizing role just mentioned that it currently performs its most important work in science. Contrary to the claims made by some of the critics mentioned previously, meta-analysis can be

regarded as a legitimate, and important, means of detecting empirical phenomena in the behavioral sciences (Gage, 1996). I briefly refer to the achievements of meta-analysis when considering the matter of scientific progress in psychology in the next section.

It is worth remarking at this point that, although meta-analysis can be considered a form of constructive replication, this should not be taken as sufficient grounds for thinking that there is not a major problem with replication in psychology. In response to the many expressions of concern that there is a dearth of replication studies in psychology, Schmidt and Oh (2016) counter that many primary research studies are in fact being replicated because they are included in meta-analyses, which are themselves replications. Instead, they think that publication bias, and questionable research practices, such as selectively reporting *p* values, are the real problem. I think that Schmidt and Oh are correct in claiming that meta-analyses do provide us with conceptual replications, a point that is seldom made. However, the major problem with their line of argument is that it discounts the importance of direct replications in science, which are undertaken to duplicate the sampling and experimental procedures of the original research to confirm their findings. Importantly, close replication is a desirable, often necessary, precursor to constructive replication (Haig, 2014). It is the paucity of direct replications in psychology that forms the basis for the claim that psychology has a replication problem.

Meta-Analysis and Scientific Explanation

Given that the detection of empirical phenomena and the construction of explanatory theories are quite different research undertakings, which generally employ different types of methods, the suggestion that meta-analysis can *directly* contribute to the construction of explanatory theory (Cook et al., 1992; Schmidt, 1993) is a surprising methodological claim. In approving this extension of meta-analysis beyond a concern with phenomena detection, Schmidt (1993) acknowledges that scientific explanation normally involves the causal explanation of observed phenomena. Nevertheless, he maintains that it is appropriate to regard scientific explanation to include "all research processes that can contribute ultimately to theory building, including the first step of

determining what the relationships are among important variables or constructs and how stable these relationships are" (p. 1164). Thus, the demonstration of a general effect, such as the pervasive influence of psychoeducational treatments on adult surgical patients, is deemed a meta-analysis at the "lowest level of explanation." On the other hand, the use of meta-analysis to test competing theories of how patients cope with the stress of surgery is viewed as higher level explanatory meta-analysis.

However, this attempt to extend the role of meta-analytic methods beyond that of phenomena detection really is something of a sleight of hand, whose semblance of plausibility derives from playing fast and loose with the relationship between phenomena detection and scientific explanation. As we have seen in the previous section of this chapter, Schmidt's (1993) "ubiquitous positive effects" are empirical phenomena, and statements about phenomena are the *objects*, or targets, of scientific explanation; they are not the explanations themselves. The question What do statements of empirical phenomena explain? wants for a sensible reply. This is not surprising because the successful detection of phenomena is essentially a descriptive achievement that involves investigative practices that are, for the most part, quite different from explanatory endeavors. Typically, the methods used in phenomena detection are statistical in kind.

By contrast, scientific explanation is often mechanistic in nature (e.g., Bechtel & Abrahamsen, 2005; Salmon, 1984). That is to say, explanation requires the identification of the causal mechanisms that underlie and give rise to empirical phenomena, along with a detailing of the ways in which those mechanisms produce the phenomena we seek to understand. Determining predictive success is, of course, a common strategy for evaluating scientific theories, and it is true that testing explanatory theories by using meta-analytic methods can provide one with a measure of the predictive success of theories. However, in this role meta-analysis is not directly concerned with their explanatory adequacy. Meta-analysis itself is not an explanatory approach to theory evaluation (Chow, 1987). To employ meta-analysis to assist in the predictive testing of an explanatory theory does not thereby confer an explanatory role on meta-analysis itself; one does not properly assign status simply on the basis of association. To repeat, when Schmidt (1993) calls phenomena detection

explanation, he conflates a methodological distinction of fundamental importance.

With meta-analysis in mind, a question worth asking at this point is, Has psychology made good progress in its quest to detect empirical phenomena? Some psychologists doubt that this is so. For example, Gergen (1973) maintains that the behavioral sciences deal with facts that are often nonrepeatable, and that, at best, these sciences produce generalizations that hold for a limited time only because they are invalidated by cultural and historical factors. Partly for this reason, he distrusts meta-analysis as a basis for claiming that empirical generalizations exist (Gergen, 1994). Relatedly, Cronbach (1975) believes that the interactive complexity of psychology's subject matter ensures that its generalizations have a short half-life. Furthermore, Lykken (1991) argues that psychology has made poor empirical and theoretical progress and, with respect to the former, contends that many of its empirical findings fail to replicate (a matter that is currently a topic of much empirical research and debate).

In the face of negative assessments such as these, Gage (1996) counters that the results of meta-analysis include an array of stable, and robust, first-order and interaction effects that support the conclusion that the behavioral sciences have detected numerous empirical phenomena worthy of theoretical explanation. Furthermore, Hedges (1987) provides an example of one type of study that is needed to make informed judgments about empirical progress in psychology. He shows that a comparison of the empirical consistency of the results of replicated exemplary experiments in physics and psychology, which use the same numerical methods, reveals a similar degree of empirical cumulation. This is a piece of knowledge about empirical progress in psychology that challenges popular opinion. Importantly, Hedges distinguishes between empirical and theoretical cumulativeness and appropriately notes that any deficiency in theory in the social and behavioral sciences would not seem to be the result of the inability of those disciplines to replicate experiments under good conditions.

As noted in Chapter 3, psychology's heavy reliance on its own interpretation of null hypothesis significance testing is an indefensible practice. Orlitzky (2012) constructively recommends a package of reforms designed to help to overcome this problem. One of these reforms involves placing greater emphasis on abductive research methods, which are concerned with explanatory

inference. Although I strongly endorse this suggestion (Haig, 2005, 2014), I think that Orlitzky's understanding of the role of abduction in a number of research methods is either unclear or confused. My primary purpose in this section of the chapter is to consider what role, if any, abduction has in meta-analysis. However, before doing so, I briefly comment on two other methods that Orlitzky believes directly involve abductive reasoning.

Orlitzky (2012) takes the methods of exploratory data analysis and computer-intensive resampling to be basically abductive in nature. Regarding exploratory data analysis, the method is, as its name implies, *data* analytic in character. It performs no explanatory role. As noted in Chapter 2, exploratory data analysis involves descriptive, and frequently quantitative, detective work designed to reveal structure, or patterns, in the data. For this reason, I do not think it can be considered an *abductive*, or explanatory, undertaking in any interesting sense of the term. In Chapter 2, I made this very point against Behrens, Dicerbo, Yel, and Levy's (2013) treatment of exploratory data analysis.

Similarly, computer-intensive resampling methods serve a data analytic purpose, not an explanatory one. They are confirmatory procedures designed to check the reality of the patterns revealed by exploratory data analysis. As such, they enable one to adopt the "just checking" strategy of close replication; they do not directly contribute to explanatory research. In this descriptive, confirmatory role, they can be seen as part of the overall process of detecting empirical phenomena. As has been emphasized at several places in this book, this process is quite different from building explanatory theories. Such theories are often introduced to understand empirical phenomena, and the type of reasoning involved in their construction is abductive in nature.

Is meta-analysis essentially an abductive method? Orlitzky thinks that it can be abductive in nature, although it need not be. I disagree. Orlitzky makes his case by taking the argument schema for existential abductive inference that I laid out in my characterization of exploratory factor analysis in Chapter 3 and instantiating it with the following meta-analytic example:

> The surprising empirical phenomenon of regularities . . . between corporate, social and financial performance is identified. If corporate social performance helps build

> corporate legitimacy and reputation . . . and thus reduces business risk, . . . and if we can have some confidence in the reliability and validly of measures of corporate social performance, . . . then the observed phenomenon would follow as a matter of course. Hence, there are grounds for judging the risk-reputation hypothesis to be initially plausible and worthy of further pursuit. (Orlitzky, 2012, p. 205)

In the second premise of this schema, Orlitzky inserts information about the causal hypothesis of risk reputation that explains the empirical phenomenon described in the first premise. However, I fail to see how the explanatory information contained in the risk-reputation hypothesis can be generated *directly* by the use of meta-analytic techniques. Instead, I believe that it is gained by abductively hypothesizing plausible causes, without directly employing the resources of meta-analysis to do so. Even so-called explanatory meta-analysis (Cook et al., 1992) contains no explicit abductive methodology. The fact of the matter is that meta-analytic techniques are suited for identifying empirical phenomena (a point Orlitzky acknowledges), whereas explanations for phenomena are fashioned abductively, with or without the help of codified abductive methods for doing so.

Conclusion

The currently popular practice of conducting meta-analytic reviews of empirical studies has been examined in respect of some of its conceptual foundations. My examination began by presenting and evaluating Glass's little-known rationale for employing meta-analysis to conduct evaluative, as opposed to scientific, research. Glass's strong distinction between scientific and evaluative inquiry was found wanting, although meta-analysis was judged to be suitable for evaluative inquiry. The major negative consequence of drawing this hard-and-fast distinction was that it prevented Glass from appreciating that meta-analysis has a major role in helping scientific researchers establish empirical phenomena. David Sohn's extended argument that meta-analysis is not a vehicle for scientific discovery was then examined. It was found wanting for a number of reasons, including Sohn's reluctance to back up key assertions about challenges to meta-analysis, such as correcting for publication bias.

The remainder of the chapter focused primarily on the role of meta-analysis in detecting empirical phenomena and constructing explanatory theories. The question of whether meta-analysis as a direct role in the construction of explanatory theories was answered in the negative. The importance of the distinction between phenomena detection and theory construction was reaffirmed, with each endeavor being shown to have quite different roles and employ different methods, in scientific inquiry. Relatedly, Orlitzky's claim that meta-analysis can be a vehicle for abductive reasoning was judged to be implausible.

More work on the philosophical foundations of meta-analysis is clearly needed, although the further reading suggested next contains conceptual work not dealt with in this chapter. However, from this selective examination of its conceptual foundations, it can be concluded that meta-analysis receives its primary justification in scientific research by articulating one important way in which researchers can fashion empirical generalization from the findings of primary studies. Its value in this role stems directly from the importance accorded the goal of phenomena detection in science.

Further Reading

Gene Glass is the lead author of the first book published on the methodology of meta-analysis (Glass, G. V., McGaw, B., & Smith, M. L., *Meta-analysis for social science*. Beverly Hills, CA: Sage, 1981).

Glass contrasts the distinction between scientific and evaluative inquiry in his (1972) article, "The Wisdom of Scientific Inquiry on Education" (*Journal of Research in Science Teaching, 9*, 3–18, 1972). In his later article with Reinhold Kliegl, he defends the basic tenets of his approach by considering a number of issues in the philosophy of science. See their "An Apology for Research Integration in the Study of Psychotherapy" (*Journal of Consulting and Clinical Psychology, 51*, 28–41, 1983).

The third edition of Frank Schmidt and John Hunter's book *Methods of Meta-analysis: Correcting Error and Bias on Research Findings* (Los Angeles, CA: Sage, 2015) represents an important, and somewhat different, approach to meta-analysis, whose initial development began at about the same time as Glass began his work on the topic.

Schmidt's view of the nature of psychological science and the important role of meta-analysis within it is adumbrated in the following two articles: "What Do Data Really Mean? Research Findings, Meta-analysis, and Cumulative Knowledge in Psychology" (*American Psychologist, 47*, 1173–1181, 1992); and, with I.-S. Oh, "The Crisis of Confidence in Research Findings in Psychology: Is

Lack of Replication the Real Problem? Or Is It Something Else?" (*Archives of Scientific Psychology, 4*, 32–37, 2016).

Michael Borenstein et al.'s *Introduction to Meta-analysis* (Chichester, England: Wiley, 2009) is a comprehensive and informative treatment of meta-analysis that stresses a conceptual understanding of the topic written by four experts on meta-analysis.

Larry Hedges has been a major contributor to the development of the statistical foundations of meta-analysis. His book with Ingram Olkin, *Statistical Methods for Meta-analysis* (Orlando, FL: Academic Press, 1985), remains the authoritative source on its statistical foundations.

An informative historical perspective on the impact and origins of meta-analysis in the social sciences is William Shadish and Jesse Lecy's, "The Meta-analytic Big Bang" (*Research Synthesis Methods, 6*, 246–264, 2014).

Hans J. Eysenck was one of the most vociferous critics of meta-analysis. See his, "Meta-analysis: An Abuse of Research Integration" (*Special Education, 18*, 41–59, 1984) for an extended critique of the methodology. Eysenck contends that meta-analysis fails to objectively determine empirical facts and that it ignores the overriding importance of theory in science.

The philosopher of science, Edward Erwin, examines a number of conceptual and epistemological issues that arise from the use of meta-analysis to evaluate the effectiveness of psychotherapy. See his, "Establishing Causal Connections; Meta-analysis and Psychotherapy" (*Midwest Studies in Philosophy, 9*, 421–436, 1984). Erwin argues that meta-analysis alone cannot solve all the problems that arise in integrating findings from multiple studies.

Another philosopher of science, Jacob Stegenga, criticizes meta-analysis for being insufficiently intersubjective in its assessments of hypotheses. Because of this lack of objectivity, he believes that the epistemic prominence given to meta-analysis is unjustified. See his, "Is Meta-analysis the Platinum Standard of Evidence?" (*Studies in History and Philosophy of Biological and Biomedical Sciences Part C, 42*, 497–507, 2011).

David Sohn's, "Meta-analysis and Science" (*Theory and Psychology, 6*, 229–246, 1995) challenges the claim that meta-analysis can make discoveries about nature. This chapter evaluates Sohn's main argument against meta-analysis.

Regan Shercliffe, William Stahl, and Megan Tuttle, in their "The Use of Meta-analysis in Psychology: Superior Vintage or the Casting of Old Wine in New Bottles?" (*Theory & Psychology, 19*, 413–430, 2009), challenge the widely held view that meta-analysis is superior to other approaches to literature reviews, such as narrative reviews. They suggest that all forms of literature review have their strengths and limitations and argue for the importance of theory in deciding which topics to review and the best approach to adopt.

References

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, 99, 388–399.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*, 421–441.

Behrens, J. T., Dicerbo, K. E., Yel, N., & Levy, R. (2013). Exploratory data analysis. In J. A. Schinka, W. F. Velicer, and I. B. Weiner (Eds.), *Handbook of psychology.* (2nd ed., Vol. 2, pp. 34–70). Hoboken, NJ: Wiley.

Bhaskar, R. (1978). *A realist theory of science* (2nd ed.). Sussex, England: Harvester Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Bruno, J. E., & Ellett, F. S. (1988). A core analysis of meta-analysis. *Quality and Quantity*, 22, 111–126.

Chow, S. L. (1987). Meta-analysis of pragmatic and theoretical research: A critique. *Journal of Psychology*, *121*, 259–271.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation: A case book.* New York, NY: Russell Sage Foundation.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, 48, 449–472.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116–127.

DeWaele, J.-P., &. Harré, R. (1979). Autobiography as a psychological method. In G. P. Ginsburg (Ed.), *Emerging strategies in social psychological research* (pp. 177–209). Chichester, England: Wiley.

Erwin, E. (1984). Establishing causal connections: Meta-analysis and psychotherapy. *Midwest Studies in Philosophy* (Vol. 9, pp. 421–436). Minneapolis, MN: University of Minnesota Press.

Eysenck, H. J. (1984). Meta-analysis: An abuse of research integration. *Special Education*, 18, 41–59.

Gage, N. L. (1996). Confronting counsels of despair for the behavioral sciences. *Educational Researcher*, *25*, 5–15, 22.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*, 309–320.

Gergen, K. J. (1994). *Toward transformation in social knowledge* (2nd ed.). Thousand Oaks, CA: Sage.

Glass, G. V. (1972). The wisdom of scientific inquiry on education. *Journal of Research in Science Teaching*, *9*, 3–18.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8.

Glass, G. V., & Kliegl, R. M. (1983). An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology*, *51*, 28–41.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gottfredson, G. D. (1984). A theory-ridden approach to programme evaluation. *American Psychologist*, *39*, 1101–1112.

Haack, S. (1974). *Deviant logic*. Cambridge, England: Cambridge University Press.

Habermas, J. (1971). *Knowledge and human interests*. Boston, MA: Beacon Press.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*, 371–388.

Haig, B. D. (2009). Detecting psychological phenomena: Taking bottom-up research seriously. *American Journal of Psychology*, *126*, 135–153.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences.* Cambridge, MA: MIT Press.

Haig, B. D., & Borsboom, D. (2012). Truth, science, and psychology. *Theory & Psychology*, *22*, 272–289.

Harré, R., & Madden, E. H. (1975). *Causal powers.* Oxford, England: Basil Blackwell.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, *42*, 443–455.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York, NY: Academic Press.

Hooker, C. A. (1987). *A realistic theory of science.* New York, NY: State University of New York Press.

Ioannides, J. P. A. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, *94*, 485–514.

Kitcher, P. (1983). *The nature of mathematical knowledge*. Oxford, England: Oxford University Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 41, 429–471.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology, Vol. 1: Matters of public interest* (pp. 3–39). Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Molenaar, P. C. M. (2004). A manifesto on psychology as an idiographic science: Bringing the person back into psychology, this time for ever. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 201–218.

Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods*, *15*, 199–228.

Prioleau, L., Murdock, M., & Brody, N. (1983). An analysis of psychotherapy versus placebo studies. *The Behavioral and Brain Sciences*, *6*, 275–310.

Quine, W. V. (1953). *From a logical point of view*. Cambridge, MA: Harvard University Press.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, *85*, 185–193.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.

Schmidt, F. L. (1993). Meta-analysis and cumulative knowledge. *Contemporary Psychology*, *38*, 1163–1165.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*, 32–37.

Shadish, W. R., & Lecy, J. D. (2014). The meta-analytic big bang. *Research Synthesis Methods*, *6*, 246–264.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher*, *13*, 6–15.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

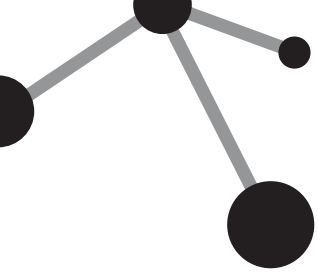
Sohn, D. (1995). Meta-analysis as a means of scientific discovery. *American Psychologist*, *50*, 108–110.

Sohn, D. (1996). Meta-analysis and science. *Theory and Psychology*, *6*, 229–246.

Sohn, D. (1997). Questions for meta-analysis. *Psychological Reports*, *81*, 3–15.

Unger, R. M. (1975). *Knowledge and politics*. New York, NY: Free Press.

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*, 393–472.



6

EXPLORATORY FACTOR ANALYSIS

Factor analysis and component analysis are two broad classes of procedures that share a common goal: to reduce a set p observed variables to a set of m new variables (m < p).

—W. F. Velicer and D. N. Jackson, 1990

Exploratory factor analysis is an abductive method for formulating hypotheses using the common cause principle, but also to be used along with confirmatory factor analysis, which tests hypotheses.

—S. A. Mulaik, 2010

Introduction

Factor analysis is an important family of multivariate statistical methods that is widely used in the behavioral and social sciences. It is also employed to some extent in the biological and physical sciences. The best known model of factor analysis is common factor analysis, which has its origins in Charles Spearman's (1904) pioneering work on the nature of general intelligence. With common factor analysis, each "observed" or manifest variable in a set of manifest variables is a linear function of one or more latent

common factors and one unique factor. There are two main types of common factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is used to investigate the underlying structure of correlations among observed or manifest variables. The goal of EFA is to describe this structure in an economical manner by hypothesizing a small number of factors or latent variables that are thought to underlie and give rise to the patterns of correlations in new domains of manifest variables. Intellectual abilities, personality traits, and social attitudes are well-known classes of latent variables that are the products of factor analytic research. EFA is in effect a method that facilitates the generation of hypotheses or theories that explain patterns of correlations. By contrast, CFA formulates theories about the latent factors and how they relate to the manifest variables and then tests their structure. CFA has become more prominent in factor analytic studies than EFA, though EFA continues to be widely used.

Despite the advanced statistical state and regular use of EFA, debate about its basic nature and worth continues. Regarding its nature, many factor analytic methodologists take EFA to be a method for hypothesizing latent variables to explain patterns of correlations. However, some take it to be a method of data reduction that provides an economical description of correlational data. Further, with the advent of CFA and full structural equation modeling, the prominence of EFA in multivariate research has declined. Today, methodologists and researchers often recommend and employ CFA as the method of choice in factor analytic studies.

The major goal of this chapter is to examine the conceptual foundations of EFA and argue for the view that it is properly construed as a method for generating rudimentary explanatory theories. In the first half of the chapter, I contend that EFA is an abductive method of theory generation that exploits an important, but underappreciated, principle of scientific inference known as the *principle of the common cause.* It is surprising that this characterization of the inferential nature of EFA rarely figures explicitly in the factor analytic literature because it contributes in an important way to its abductive nature, and it coheres well with the generally accepted view of EFA as a latent variable method. In the second half of the chapter, I discuss a number of additional methodological issues that arise in critical discussions of EFA. In particular, I argue that the principle of the common cause supports a realist,

not an instrumentalist, interpretation of factors; that factorial theories have genuine, albeit modest, explanatory merit; that the methodological challenge of factor indeterminacy can be satisfactorily met by EFA; that EFA can in fact discover causes; that EFA is quite different from principal components analysis (PCA); and that as a useful method of theory generation, EFA can be profitably employed in tandem with CFA and other methods of theory evaluation. I conclude that, if understood and used properly, EFA serves as a useful generator of rudimentary explanatory theories.

Exploratory Factor Analysis and Scientific Inference

The Nature of Abductive Inference

Abduction is a form of reasoning involved in the generation and evaluation of *explanatory* hypotheses and theories. In recent decades, developments in the fields of philosophy of science, artificial intelligence, and cognitive science (e.g., Josephson & Josephson, 1994; Magnani, 2001; Thagard, 1988, 1992) have significantly advanced our understanding of abductive reasoning. It is now known that there are a number of different ways in which explanatory hypotheses can be abductively obtained. For example, in focusing on the generation of hypotheses, Thagard (1988) helpfully distinguishes between *existential abduction*, which hypothesizes the existence of previously unknown objects or properties, and *analogical abduction*, which employs successful past cases of hypothesis generation to form new hypotheses similar to relevant existing ones. The next section suggests that existential abduction is the type of abduction involved in the factor analytic production of explanatory hypotheses, although analogical abduction also is sometimes employed in this regard.

It is common for philosophers to characterize abduction in terms of the logical form of an argument. This can be done as follows:

> The surprising empirical phenomenon, P, is detected.
> But if hypothesis H were approximately true, and the relevant auxiliary knowledge A were invoked, then P would follow as a matter of course.
> Hence, there are grounds for judging H to be initially plausible and worthy of further pursuit.

This characterization of an abductive argument accommodates the following important features of science: It is typically empirical phenomena, not data, that hypotheses are produced to explain, the role of background knowledge is needed for the derivation of hypotheses, a regulative role should be assigned to truth, and initial plausibility assessments feature centrally in the generation and development of new knowledge.

Exploratory Factor Analysis and Abductive Inference

We turn now to consider the claim that EFA is most fundamentally an abductive method of theory generation. As noted, existential abductions in science often hypothesize the existence of entities previously unknown to us. The innumerable examples of existential abduction in science include the initial postulation of entities such as atoms, phlogiston, viruses, tectonic plates, Spearman's *g*, habit strength, and extraversion. We now know that some of these entities exist, that some of them do not exist, and that we are unsure about the existence of others. In cases like these, the initial abductive inferences are made to claims primarily about the *existence* of theoretical entities to explain empirical facts or phenomena. Thus, in the first instance, the hypotheses given to us through the use of EFA do little more than postulate the existence of the latent variables in question. They say little about their nature and function, and it remains for further research to elaborate on the first rudimentary conception of these variables.

The factor analytic use of existential abduction to infer the existence of the theoretical entity *g* can be coarsely reconstructed in accord with the previous schema for abductive inference along the following lines:

> The surprising empirical phenomenon known as the *positive manifold* is identified.
> If *g* exists, and it is validly and reliably measured by a Wechsler intelligence scale (or some other objective test), then the positive manifold would follow as a matter of course.
> Hence, there are grounds for judging the hypothesis of *g* to be initially plausible and worthy of further pursuit.

It was remarked previously that the factor analytic generation of hypotheses is sometimes a mixture of existential and analogical

abduction, where we simultaneously posit the existence of a latent variable and offer the beginnings of a characterization of that entity by brief analogy to something that we understand quite well. Recall that analogical abduction appeals to known instances of successful abductive hypothesis formation to generate new hypotheses like them.

To accommodate the presence of analogical abduction, the abductive argument schema just given would need an additional premise that indicates there is reason to believe that a hypothesis of the appropriate kind would explain the positive manifold. When Charles Spearman first posited general intelligence to explain correlated performance indicators, he thought of it as mental energy, likening it to physical energy—a process well understood by the physics of the time. His initial inference to claims about *g*, then, was a blend of existential and analogical abduction.

This example serves to illustrate the point that methodologists should take the method of EFA proper to include the factor analyst's substantive interpretation of the statistical factors. In this regard, it is important to realize that the exploratory factor analyst has to resort to his or her own abductive powers when reasoning from correlational data patterns to underlying common causes. This point can be brought out by noting that the schema for abduction, and its application to the factor analytic generation of Spearman's hypothesis of *g*, are concerned with the form of the arguments involved and not with the actual generation of the explanatory hypotheses. In each case, the explanatory hypothesis is *given* in the second premise of the argument. An account of the genesis of the explanatory hypothesis must therefore be furnished by some other means. It is plausible to suggest that reasoning to explanatory hypotheses trades on our evolved cognitive ability to abductively generate such hypotheses. The modern originator of abductive inference, Charles Peirce, maintains that the human ability to engage readily in abductive reasoning was founded on a guessing instinct that has its origins in evolution. More suggestively, Carruthers (2002) claims that our ability to engage in explanatory inference is almost certainly largely innate, and he speculates that it may be an adaptation selected for because of its crucial role in the fitness-enhancing activities of our ancestors, such as hunting and tracking. Whatever its origin, an informative methodological characterization of the abductive nature of factor analytic inference must appeal to the scientist's own psychological resources as well as those of logic. It is a

tenet of realist methodology that a full characterization of knowledge production must make reference to the knowing subject.

Before leaving consideration of the general abductive nature of EFA, it should be briefly noted that there are a number of special features of EFA that play an important role in facilitating the abductive generation of hypotheses. To take one example, simplicity, or parsimony, is an important desideratum in fashioning scientific explanations, and Thurstone's (1947) criteria for simple structure combine in an explicit formulation of parsimony in EFA. Stated in the distinctive language of factor analysis, Thurstone's insight was to appreciate that rotation to the oblique simple structure solution provided an objective basis for acceptable terminal factor solutions that included reference to latent as well as manifest variables.

The Principle of the Common Cause

It is now time to consider the important methodological principle that drives and shapes the nature of the existential abductive inference involved in EFA. It is well known that EFA is a common factor analytic model in which the latent factors it postulates are referred to as *common* factors. Not surprisingly, these factors are often understood, and sometimes referred to, as common *causes.* Yet, seldom have factor analytic methodologists attempted to formulate a principle, or maxim, of inference that guides the reasoning to common causes. There is, however, an important principle of scientific inference, known in philosophy of science as the *principle of the common cause*, that can be used to good effect here. In what follows, I discuss the principle of the common cause and then spell out its central role in EFA. This principle drives and shapes the nature of the existential abductive inference involved in EFA.

The principle of the common cause has received some consideration in the philosophical literature and occasionally appears to be tacitly employed in behavioral research. However, it has been widely ignored in general scientific methodology. In explicitly introducing the idea of the principle of the common cause, Hans Reichenbach (1956) was concerned to capture the idea that if two events, A and B, are correlated, then one might be the cause of the other. Alternatively, they might have a common cause C, where this cause always occurs before the correlated events. Reichenbach was the first to make this idea precise, and he did so by formulating it as a statistical problem.

He suggests that the common cause C is said to "screen off" the correlation between A and B, when A and B are uncorrelated, conditional upon C. A common cause screens off each effect from the other by rendering its correlated effects (conditionally) probabilistically independent of each other. For example, given the occurrence of a flash of lightning in the sky, a correlation between two people apparently observing that flash not only is a coincidence, but also is due to the flash of lightning being a common cause. Further, the probability of one person seeing the flash of lightning, given that it does occur, is not affected by whether the other person observes the lightning flash. Reichenbach's principle of the common cause can thus be formulated succinctly as follows: "Simultaneous correlated events have a prior common cause that screens off the correlation."

Later work (Arntzenius, 1993; Salmon, 1984; Sober, 1988) suggests that Reichenbach's formulation of the principle needs to be amended in a number of ways. First, not every improbable coincidence, or significant correlation, has to be explained through a common cause. For this reason, the principle is sometimes taken to say, "If an improbable co-incidence has occurred, and there is no direct causal connection between the coincident variables, then one should infer a common cause." However, this amendment does not go far enough, for there are a number of other possible alternative causal interpretations of correlations. For example, two correlated variables might be mediated by an intervening cause in a developmental sequence, or they might be the result of separate direct causes, and so on. Plausible inference to a common cause must rule out alternative causal interpretations like these. We may, therefore, further amend Reichenbach's formulation of the principle as follows: "Whenever two events are improbably, or significantly, correlated, we should infer a common cause unless we have good reason not to." Clearly, the principle should not be taken as a hard-and-fast rule, for, in many cases, proper inferences about correlated events will not be of the common causal kind. The qualifier "unless we have a good reason not to" should be understood as an injunction to consider causal interpretations of the correlated events other than common causes. Also, there will be occasions when it is incorrect to draw any sort of causal conclusion. Some correlations are accidental correlations that are not brought about by causes.

The existence of different attempts to improve on Reichenbach's (1956) initial formulation of the principle of the common cause

suggests that there might be more than one acceptable version of the principle. We might expect this to be the case because different subject matters in different domains might well require different formulations of the principle. For example, Reichenbach, a philosopher of physics, took the principle to apply to correlated events that are spatially separated. However, social and behavioral scientists regularly infer common causes for events that are not spatially separated. This is clearly the case if the correlated variables are performance measures on tests of intelligence and personality. Further, Sober (1988) argues that in evolutionary theory, phylogenetic inference to common ancestry involves postulating a common cause, but this will be legitimate only if certain assumptions about the process of evolution are true. Thus, in formulating a principle of the common cause in a way that can be used effectively in a given domain, relevant contingent knowledge about that domain will shape the formulation of the principle and moderate its use. Routine use of a fixed, general formulation of the principle of the common cause that reasons from correlational data alone is unlikely to lead consistently to appropriate conclusions.

Two related features of the principle of the common cause should also be acknowledged: As Salmon (1984) has observed, the principle is sometimes used as a principle of explanation (we appeal to common causes to *explain* their correlated effects), and it is sometimes used as a principle of inference (we use the principle to *reason* to common causes from their correlated effects). The principle of the common cause is a form of abductive inference where one reasons from correlated events to common causes thought to explain those correlations. Thus, we should go further than Salmon and claim that the principle of the common cause simultaneously combines these explanatory and inferential features to yield explanatory inferences.

Exploratory Factor Analysis and the Principle of the Common Cause

It is sometimes said that the central idea in factor analysis is that the relations between a large number of observed variables are the direct result of a smaller number of latent variables. McArdle (1996) maintains that this is a theoretical principle employed in empirical research to identify a set of underlying factors. However,

although true of EFA, this principle does not constrain factor analysts to infer the *common* latent factors that are the appropriate outcome of using common factor analysis. For this to happen, the principle has to be linked to the principle of the common cause or recast in more specific methodological terms in accordance with that principle. The principle of the common cause not only directs one to infer common causes, but also it assumes that those inferences will be to relatively few common causes. Reichenbach's (1956) original formulation of the principle, which allows inference to just one common cause, is obviously too restrictive for use in multiple factor analysis. However, amending the principle to allow for more than one common cause, combined with the restraint imposed by following Ockham's razor (do not multiply entities beyond necessity), will enable one to infer multiple common causes without excess.

Although EFA is used to infer common causes, expositions of common factor analysis that explicitly acknowledge the importance of the principle of the common cause are difficult to find. Kim and Mueller's (1978) basic exposition of factor analysis is a noteworthy exception. In discussing the conceptual foundations of factor analysis, these authors evince the need to rely on what they call *the postulate of factorial causation*. The postulate of factorial causation is characterized by them as "the assumption that the observed variables are linear combinations of underlying factors, and that the covariation between observed variables is solely due to their common sharing of one or more of the common factors" (p. 78). The authors make clear that the common factors mentioned in the assumption are to be regarded as underlying causal variables. Taken as a methodological injunction, this postulate functions as a variant of the principle of the common cause. Without appeal to this principle, factor analysts could not identify the underlying factor pattern from the observed covariance structure. It should be noted, further, that in his examination of the philosophical foundations of factor analysis, Yu (2006) explicitly discusses the important role of the principle of the common cause in the method.

Two features of the principle of the common cause that make it suitable for EFA are that it can be applied if we do not know how *likely* it is that the correlated effects are due to a common cause (this feature is consistent with the views of Reichenbach [1956], Salmon

[1984], and Sober [1988] on common causal reasoning) and also in situations where we are essentially ignorant of the *nature* of the common cause. The abductive inference to common causes is a basic explanatory move that is nonprobabilistic, and qualitative, in nature. It is judgments about the soundness of the abductive inferences, not the assignment of probabilities, that confer initial plausibility on the factorial hypotheses spawned by EFA.

It is important to appreciate that the principle of the common cause does not function in isolation from other methodological constraints. Embedded in EFA, the principle helps to limit existential abductive inference to those situations where we reason back from *correlated* effects to one or more *common* causes. Although covariation is an important basic datum in science, not all effects are expressed as correlations, and, as noted previously, not all causes are of the common causal variety. It follows from this that researchers should not always expect common causal interpretations of multivariate data for there are numerous alternative latent variable models. The simplex model of latent variables is a case in point (e.g., Mulaik & Millsap, 2000). Further, the frequency of the proper use of EFA should be much less than the frequency of proper use of the principle of the common cause because the principle can be employed by non–factor analytic means.

In this first half of the chapter, it was argued that an appeal to abductive inference, linked to the principle of the common cause, leads naturally to the view that EFA is an abductive method of theory generation that enables researchers to theorize the existence of latent variables. Although this method uses the statistical ideas of multiple regression and partial correlation, it does so to facilitate inferences to the latent variables. On the view presented here, EFA is glossed as a set of multivariate procedures that help us reason in an existentially abductive manner from robust correlational data patterns to plausible explanatory prototheories via the principle of the common cause.

Fictionalism and the Principle of the Common Cause

In evaluating factor analytic arguments for general intelligence, Ned Block (1976) identifies, and critically examines, an assumption of factor analysis that he calls the *correlation-entails-commonality principle*. He states the principle informally in these terms: "Insofar

as two tests of abilities correlate, this correlation is totally due to a common ability measured by both tests" (p. 129). This principle is reasonably interpreted as a naïve version of the principle of the common cause. However, my purpose here is not to examine the naiveté of the principle, but to consider Block's surprising suggestion that the principle is supported by the doctrine of fictionalism. Fictionalism is the part of an instrumentalist view of theories that maintains that theoretical terms in science such as *electron, gene,* and *g* should not be taken as referring to unobserved entities because entities such as these do not exist. Instrumentalism is the antirealist doctrine that scientific theories are neither true nor false, but are more or less useful devices for the summary and prediction of empirical relationships.

Applied to EFA, fictionalism dictates that factor constructs do not refer to underlying latent causes viewed as theoretical entities. Rather, they are summary expressions of the way manifest variables covary. However, if fictionalism supports the correlation-entails-commonality principle, as Block suggests, and a version of this principle is central to factor analysis, as has been argued, then factor analysis cannot be used by factor analytic researchers to make inferences about latent variables. Does this mean, then, that ability and trait theorists of realist persuasion such as Spearman, Thurstone, Cattell, and Costa and Macrae have misused common factor analysis in this way? I do not think so for, as a variant of the principle of the common cause, the correlation-entails-commonality principle should be understood as a scientific principle that sanctions inference to common causes, *wherever they may lie.* In psychology, most of our claims about common causes are the result of an explanatory strategy that appeals to latent variables that are thought to reside within the organism. The principled empiricist might well look for manifest common causes in the environment, but the disadvantage of his or her philosophy is that it allows this empiricist to look for them only there. The realist, by contrast, can invoke the principle of the common cause without such ontological restriction and posit latent common causes if they are thought to reside within the organism. Not only does a realist's use of the principle of the common cause enable factor analysts to extend their referential reach to latent variables, it also bestows a measure of credibility on the associated inferences.

The second half of the chapter defends the realist interpretation of EFA presented thus far. As noted in the chapter's introduction,

this is done by attending to a number of methodological criticisms that have been made against the method.

Methodological Challenges to Exploratory Factor Analysis

The Explanatory Merit of Factorial Theories

One challenge to the interpretation of EFA as an abductive method of theory generation is the claim that the theories it produces are of little explanatory worth. In countering this criticism, I suggest that factorial theories spawned by EFA are essentially dispositional in nature, and that dispositional theories do have genuine, though limited, explanatory import (Rozeboom, 1984; Sober, 1982). Existential abduction, it will be recalled, postulates the existence of new entities without being able to characterize their nature. Thus, in exploiting this form of abduction, EFA provides us with an essentially dispositional characterization of the latent entities it postulates.

Dispositional theories provide us with oblique characterizations of the properties we attribute to things by way of their presumed effects under specified conditions (Mumford, 1998; Tuomela, 1978). For example, the brittleness of glass is a dispositional property causally responsible for the breaking of glass objects when they are struck with sufficient force. Our indirect characterization of this latent property, brittleness, is in terms of the relevant striking and breaking events. Similarly, Spearman's original theory of *g* was basically dispositional in nature, for *g* was characterized obliquely in terms of children's school performance under the appropriate test conditions.

As noted immediately in the previous material, dispositional theories have often been regarded as explanatorily suspect. Perhaps the best known, and most frequently cited, example of this is Molière's scoff at explaining the soporific effects of opium by appeal to its dormitive power. However, as Rozeboom (1973) maintains, "The *virtus dormitiva* of opium *is* why people who partake of this particular substance become drowsy. Of course, that by itself leaves a great deal unknown about this power's nature, but learning of its existence and how to diagnose its presence/absence

in particular cases is a necessary preliminary to pursuit of that knowledge" (p. 67).

Similarly with EFA, the existential abductions to latent factors postulate the existence of these factors without being able to say much, if anything, about their actual nature. It is the job of EFA to help us formulate factorial hypotheses and theories about the existence of those factors, not to develop our understanding of their nature. The latter task is undertaken through the use of analogical modeling strategies. To expect EFA to develop theories, as well as generate them, is to fail to understand its proper role as a generator of dispositional theories.

An answer to the question of whether dispositional theories are of genuine explanatory worth requires us to focus on whether such theories have explanatory power. Two aspects of explanatory power that are relevant here are explanatory depth and explanatory breadth. For factorial theories, explanatory depth is naturally understood as existential depth. Existential depth is accorded those explanatory theories in science that are deep-structural in nature. Theories of this sort postulate theoretical entities that are different in kind, and hidden, from the empirical regularities they are invoked to explain. In postulating theoretical entities, deep-structural theories extend our referential reach to new entities and thereby increase the potential scope of our knowledge. The factorial theories afforded us by EFA have existential depth because the typical products of factor analytic abductions are new claims about hidden causal entities that are thought to exist distinct from their manifest effects. Existential depth deserves to be considered as an explanatory virtue of EFA's postulational theories.

The other feature of explanatory power, explanatory breadth, is a long-standing criterion of a theory's worth. Sometimes, explanatory breadth is understood as *consilience*, which is often portrayed as the idea that a theory explains more of the evidence (a greater number of facts) than its competitors. The rudimentary theories of EFA do not have consilience in this sense, for they typically do not explain a range of facts. And, they are not immediately placed in competition with rival theories. However, factorial theories of this kind are consilient in the sense that they explain the *concurrences* embodied in the relevant patterns of correlations. By appealing to common causes, these factorial theories unify their concurrences

and thereby provide us with the beginnings of an understanding of why they concur.

The two criteria that comprise explanatory power are not the only dimensions of theory appraisal that should be considered when submitting a factorial theory to preliminary evaluation. The fertility of a theory is also an important evaluative consideration. In general terms, this dimension focuses on the extent to which a theory stimulates further positive research. It should be noted here that although our initial dispositional descriptions of latent factors are low in informational content, they do not, or need not, act as a heuristic block to further inquiry as some commentators on factor analysis suggest. David Lykken (1971), for example, judges latent variable explanations from factor analysis to be "stillborn," whereas B. F. Skinner (1953) declares that they give us false assurances about the state of our knowledge. However, given that EFA trades in existential abductions, the dispositional ascription of latent factors should serve a positive heuristic function. Considered as a preliminary to what it is hoped will eventually be full-blooded explanations, dispositional ascriptions serve to define the scope of, and mark a point of departure for, appropriate research programs. Viewed in this developmental light, dispositional explanations are inquiry promoting, not inquiry blocking.

The Problem of Underdetermination

The methodological literature on factor analysis has given considerable attention to the indeterminacy of factors in the common factor model. Factor indeterminacy arises from the fact that the common factors are not uniquely determined by their related manifest variables. As a consequence, a number of different common factors can be produced to fit the same pattern of correlations in the manifest variables.

Although typically ignored by factor analytic researchers, factor indeterminacy is an epistemic fact of life that continues to challenge factor analytic methodologists. Some methodologists regard factor indeterminacy as a serious problem for common factor analysis and recommend the use of alternative methods, such as component analysis methods, because they are considered to be determinate methods. Others have countered variously that component analysis models are not causal models (and therefore are

not proper alternatives to common factor models), that they do not typically remain invariant under the addition of new variables, and that the indeterminacy of factor scores is seldom a problem in interpreting common factor analytic results because factor scores do not have to be computed.

One constructive perspective on the issue of factor indeterminacy has been offered by Mulaik and McDonald (McDonald & Mulaik, 1979; Mulaik, 1987; Mulaik & McDonald, 1978). Their position is that the indeterminacy involved in interpreting the common factors in EFA is just a special case of the general indeterminacy of theory by empirical evidence widely encountered in science, and it should not, therefore, be seen as a debilitating feature that forces us to give up on common factor analysis. Essentially, I agree with this outlook on the factor indeterminacy issue and discuss it in this light. I argue that EFA helps us produce theories that are underdetermined by the relevant evidence, and that the methodological challenge that this presents can be met in an acceptable way.

Indeterminacy is pervasive in science. It occurs in semantic, metaphysical, and epistemological forms (McMullin, 1995). Factor indeterminacy is essentially epistemological in nature. The basic idea of epistemological, or more precisely, methodological, indeterminacy is that the truth or falsity (better, acceptance or rejection) of a hypothesis or theory is not determined by the relevant evidence (Duhem, 1954). In effect, methodological indeterminacy arises from our inability to justify accepting one theory among alternatives on the basis of empirical evidence alone. This problem is sometimes referred to as the *underdetermination of theory* by *data* and sometimes as the underdetermination of theory by *evidence*. However, because theories are often underdetermined by evidential statements about phenomena, rather than data, and because evidence in theory appraisal will often be superempirical as well as empirical in nature, I refer to the indeterminacy here as the *underdetermination of theory by empirical evidence* (UTEE). Construing factor indeterminacy as a variant of UTEE is to regard it as a serious problem, for UTEE is a strong form of underdetermination that needs to be reckoned with in science. Indeed, as an unavoidable fact of scientific life, UTEE presents a major challenge for scientific methodology.

Mulaik (1987) sees UTEE in EFA as involving inductive generalizations that go beyond the data. I believe that the *inductive* UTEE

should be seen as applying specifically to the task of establishing factorial invariance, where one seeks constructive or external replication of factor patterns. However, for EFA there is also a need to acknowledge and deal with the *abductive* UTEE involved in the generation of explanatory factorial theories. The sound abductive generation of hypotheses is essentially educated guesswork. Thus, drawing from background knowledge, and constrained by correlational empirical evidence, the use of EFA can at best only be expected to yield a plurality of factorial hypotheses or theories that are thought to be in competition. This contrasts strongly with the unrealistic expectation held by many early users of EFA that the method would deliver them strongly justified claims about the one best factorial hypothesis or theory.

How, then, can EFA deal with the specter of UTEE in the context of theory generation? The answer, I think, is that EFA narrows down the space of a potential infinity of candidate theories to a manageable subset by facilitating judgments of *initial plausibility*. It seems clear enough that scientists often make judgments about the initial plausibility of the explanatory hypotheses and theories that they generate. Judgments of the initial plausibility of theories are judgments about the soundness of the abductive arguments employed in generating those theories. I suspect that those who employ EFA as an abductive method of theory generation often make compressed judgments of initial plausibility. Initial plausibility may be viewed as a constraint-satisfaction problem. Multiple constraints from background knowledge (e.g., the coherence of the proposed theory with relevant and reliable background knowledge); methodology (centrally, the employment of EFA on appropriate methodological grounds; Fabrigar, Wegener, MacCallum, & Strahan, 1999); and explanatory demands (e.g., the ability of factorial theories to explain the relevant facts in an appropriate manner) combine to provide a composite judgment of a theory's initial plausibility.

By conferring judgments of initial plausibility on the theories it spawns, EFA deems them worthy of further pursuit, whereupon it remains for the factorial theories to be further developed and evaluated, perhaps through the use of CFA. It should be emphasized that using EFA to facilitate judgments about the initial plausibility of hypotheses will still leave the domains being investigated in a state of considerable theoretical

underdetermination. It should also be stressed that the resulting plurality of competing theories is entirely to be expected and should not be thought of as an undesirable consequence of employing EFA. To the contrary, it is essential for the growth of scientific knowledge that we promote theoretical pluralism. The reason for this rests with our makeup as cognizers: We begin in ignorance, so to speak, and have at our disposal limited sensory equipment. However, we are able to develop a rich imagination and considerable powers of criticism.

These four features operate such that the only means available to us for advancing knowledge is to construct and evaluate theories through their constant critical interplay. In this way, the strategy of theoretical pluralism is forced on us (Hooker, 1987). Thus, it is through the simultaneous pursuit of multiple theories with the intent of eventually adjudicating between a reduced subset of these that one arrives at judgments of best theory.

It has been suggested that factor indeterminacy is a special case of the pervasive problem of UTEE. It has also been argued that, if we adopt realistic expectations about what EFA can deliver as a method of theory generation and also grant that the method contributes to the needed strategy of theoretical pluralism, then we may reasonably conclude that EFA satisfactorily meets this particular challenge of indeterminacy.

Can Exploratory Factor Analysis Discover Common Causes?

An important question about the worth of EFA still remains, a question that may be more important than worries about the indeterminacies of EFA: Is EFA effective enough in unearthing the common causes it hypothesizes to exist behind the correlated manifest variables? An answer to this question lies at the heart of my defense of the method. I maintain that if EFA proves to be a useful method of generating hypotheses about common causes, then worries about the various sorts of underdetermination to be found in EFA are not too unsettling for the method. There are two ways of answering this question. One is to take research programs of theory construction that make heavy use of EFA and show that the method contributes to the theoretical progress of those programs. We might want to ask, for example, whether the Spearman-Jensen theory of general intelligence is a progressive research program, or

whether the five-factor personality theory of Costa and McCrae is currently progressive. This approach would require detailed analyses of the relevant case histories, employing notions of theoretical progress that were, or are, appropriate to both science generally (a contested matter) and factor analysis more specifically. Space limitations preclude beginning such a task here, and I confine my attention briefly to the second strategy, which is to ascertain whether EFA is successful at dimensional recovery as revealed through simulations on artificial data sets where the dimensions are known in advance.

The simulation studies carried out to assess the reliability of EFA in dimensional recovery give mixed results. Some support the utility of the method, while others show poor dimensional recovery. Consider Scott Armstrong's (1967) influential, and widely cited, study, which questions the utility of EFA as a method of theory generation: Armstrong analyzed a set of artificial data in a hypothetical scenario where the underlying factors were known, and he concluded from the analysis that EFA did a poor job of recovering the known factor structure. From this, he recommended that EFA should not be used to generate theories (subsequently, many authors have cited Armstrong's article as grounds for using CFA rather than EFA in factor analytic research).

However, Preacher and MacCallum (2003) argue, correctly in my view, that Armstrong's (1967) article represents a poor piece of factor analytic research that gives misleading results, and that it provides no real basis for casting doubt on the worth of EFA as a method of theory generation. Preacher and MacCallum's study first replicated Armstrong's factor analysis on an analogous set of data and obtained essentially the same results. They then conducted a further factor analysis of that data set substituting correct factor analytic procedure for the faulty procedure used by Armstrong. Among other things, this involved using common EFA rather than PCA (strictly speaking, principal components is not a method of factor analysis), determining the correct number of factors to retain using appropriate multiple methods (the scree test and parallel analysis) and using oblique direct quartimin rotation to simple structure rather than orthogonal varimax rotation. On the basis of the congruence between the obtained factor pattern and the known structure, Preacher and MacCallum conclude that the proper use of EFA does identify the number and nature of

latent variables responsible for the manifest variables. Their exemplary use of EFA, and the well-conducted earlier simulations by factor analysts such as Thurstone and Cattell, provides good support for the view that EFA is quite good at dimensional recovery. Admittedly, these simulations dealt with simple physical systems, but Sokal, Rohlf, and Zang (1980) have shown that EFA can isolate, and help identify, meaningful biological factors that lie behind correlated physiology-of-exercise variables. The findings from good simulation studies like these, combined with the findings of a variety of empirical studies on other aspects of EFA's functioning (see Fabrigar et al., 1999), suggest that EFA can be employed as a useful generator of elementary plausible theories about common causes.

Exploratory Factor Analysis and Other Factor Analytic Methods

In this penultimate section of the chapter, I round out my characterization of EFA by briefly considering its worth in relation to the methods of PCA and CFA—two methods that are generally included in the family of factor analytic methods. In advocating the use of these last two methods, the methodological literature has sometimes argued that EFA is problematic, and that it should have a lesser role in multivariate research.

Exploratory Factor Analysis and Principal Components Analysis

Both EFA and PCA have been in existence for more than 100 years. For much of that time there have been debates about their exact nature and their relationship to each other. While Harold Hotelling's (1933) seminal formulation of PCA has its origins in the ideas of EFA, it was seen by him to be appreciably different in character. In fact, Hotelling introduced the term *components* to avoid confusion with the term *factor* in factor analysis. Today, many factor analytic practitioners regard PCA as a special case of factor analysis and employ it in preference to EFA, even though the methodological literature continues to debate whether one should do so. For example, Velicer and Jackson (1990) comprehensively reviewed a number of issues that are relevant when selecting between the two procedures and concluded that PCA should be the preferred

method for doing factor analysis. Some commentators on their article support this conclusion; others disagree with it.

The abductive view of EFA presented in this chapter endorses the claim that EFA and PCA should be regarded as different types of method (see, e.g., Fabrigar et al., 1999; Jolliff, 2002; Mulaik, 1987). Although the immediate goal of EFA is to seek recurrent data patterns through data reduction, its end goal is to identify latent variables that explain the data patterns. By contrast, PCA is a method of data reduction only. Reducing data and constructing explanations are different sorts of undertakings, and the two methods should be judged in respect of the different goals they properly serve, not in terms of their comparative efficiency in meeting a shared research goal. Further, while both EFA and PCA aim to reduce the dimensionality of data sets, they express different senses of dimensional reduction and use different techniques to achieve their goals. EFA is a causal model, with common causal structure, in which the reduced dimensions are unmeasured latent variables that are not determined by linear functions of the manifest variables. As common causes, these latent variables are arrived at abductively and serve to explain the manifest variables. By contrast, PCA assumes no explicit model and its reduced dimensions are composites of manifest variables that are determined uniquely by linear functions of the original manifest variables. Their purpose is to stand for the original manifest variables, but as such, they are statistical entities, not inferred causes, and it makes no sense to use them to try to explain the variables from which they are derived.

Because of these basic differences between the two methods, using PCA where EFA properly applies, or specifically taking the first unrotated component as a surrogate for the underlying latent variable as is sometimes suggested (e.g., Goldberg & Digman, 1994), is a cavalier ontological attitude that has serious negative consequences. Principal components are manifest variables that are not analyzed with respect to their causes, while common factors are latent variables thought to be the causes of the manifest variables from which they are derived. Thus, to take principal components as substitutes for common causes flagrantly violates the common causal presupposition on which the correct application of EFA depends. As noted in the previous section, Preacher and MacCallum (2003) demonstrated that although the proper use of

EFA can successfully identify latent variables, the use of a principal components model in its place fails to give meaningful factor analytic results.

Exploratory Factor Analysis and Confirmatory Factor Analysis

Having argued that EFA is a method that facilitates the abductive generation of rudimentary explanatory theories, it remains to consider what implications this view of EFA has for the conduct of EFA research, including its relation to the more frequently employed CFA.

The abductive view of EFA does highlight, and stress the importance of, some features of its best use, and four of these are noted. First, it should now be clear that an abductive interpretation of EFA reinforces the view that it is best regarded as a latent variable method, thus distancing it from the data reduction method of PCA. From this, it obviously follows that EFA should always be used in preference to PCA when the underlying common causal structure of a domain is being investigated.

Second, strictly speaking, the abductive interpretation of EFA also acknowledges the twin roles of the method of searching for inductive generalizations and their explanations. It should be appreciated that these research goals are different, although they are both important. It is because the detection of phenomena requires the researcher to reason inductively to empirical regularities that the abductive use of EFA insists on initially securing the invariance of factors across different populations. And, it is because the inductive regularities require explanation that one then abductively postulates factorial hypotheses about common causes.

Third, as noted previously, the abductive view of EFA places heavy emphasis on the importance of background knowledge in EFA research. In this regard, the initial variable selection process, so rightly emphasized by Thurstone (1947) and Cattell (1978), is of sufficient importance that it should be considered as part of the first step in carrying out an EFA study. For instance, in selecting the variables for his factor analytic studies of personality, Cattell was at pains to formulate and follow principles of representative sampling from a broad formulation of the domain in question. Further, the importance of background knowledge in making abductive inferences to underlying factors should not be overlooked. In this

regard, the schematic depiction of abductive inference presented previously explicitly acknowledged some of the manifold ways in which such inference depends on background knowledge. It is an important truism that the factorial hypotheses generated through abductive inference are not created ex nihilo, but come from the extant theoretical framework and knowledge of the factor analytic researcher. For most of our EFA theorizing, this source is a mix of our common sense and scientific psychological knowledge.

Finally, and relatedly, it should be made clear that acknowledging the importance of background knowledge in abductive EFA does not provide good grounds for adopting a general strategy where one discards EFA, formulates theories a priori, and uses factor analysis only in its confirmatory mode. This holds, even though when using EFA one anticipates possible common factors in order to select sufficient indicator variables to allow one to overdetermine those factors. EFA has a legitimate, indeed important, place in factor analytic research because it helpfully contributes to theory generation in at least three ways: It contributes to detection of the empirical phenomena that motivate the need for generating factorial hypotheses; it serves to winnow out a lot of theoretically possible hypotheses at the hypothesis generation stage of inquiry; and it helps to present factorial hypotheses in a form suitable for subsequent testing by CFA.

This last remark, which supports the idea that there is a useful role for abductive EFA in factor analytic research, raises the question of how EFA relates to CFA. In contrast to popular versions of the classical inductivist view of science that inductive method can generate secure knowledge claims, the use of EFA as an abductive method of theory generation can only furnish researchers with a weak logic of discovery that gives them educated guesses about underlying causal factors. It is for this reason that those who use EFA to generate theories need to supplement their generative assessments of the initial plausibility of those theories with additional consequentialist justification in the form of CFA testing or some alternative approach to theory appraisal.

In stressing the need for the additional evaluation of theories that are obtained through EFA, it should not be implied that researchers should always, or even standardly, employ classical EFA and follow this with CFA. CFA is just one of a number of options with which researchers might provide a justification of factorial hypotheses. As

an alternative, one might, for example, adopt Rozeboom's nonclassical form of EFA as a method to generate a number of models that are equivalent with respect to their simple structure by using his versatile Hyball program (1991a, 1991b) before going on to adjudicate between these models by employing CFA. Another legitimate strategy might involve formulating a causal model using EFA and following it with a procedure like that defended by Mulaik and Millsap (2000), in which a nested sequence of steps designed to test various aspects of a structural equation model is undertaken.

A further possibility, which has not been explored in the factor analytic literature, would be to follow up on the preliminary acceptance of rudimentary theories spawned by EFA by developing a number of factorial theories through whatever modeling procedures seem appropriate and then submitting those theories to a non–factor analytic form of theory appraisal. For example, it would be quite possible for competing research programs to develop theories given to them through EFA and then submit those theories to comparative appraisal in respect of their explanatory coherence. Thagard's (1992) theory of explanatory coherence, described in Chapter 4, is an integrated multicriterial method of theory appraisal that accepts as better those explanatory theories that have greater explanatory breadth, are simpler than their rivals, and are analogous to theories that have themselves been successful. This strategy of using EFA to abductively generate explanatory theories, and employing the theory of explanatory coherence in subsequent appraisals of these explanatory theories, is abductive both fore and aft.

Conclusion

Despite the fact that EFA has been frequently employed in psychological research, the extant methodological literature on factor analysis seldom acknowledges the explanatory and ontological import of the method's inferential nature. Arguably, abduction is science's chief form of creative reasoning, and the principle of the common cause is a maxim of scientific inference with important application in research. By incorporating these two related elements into its fold, EFA is ensured an important, albeit circumscribed, role in the construction of explanatory theories in psychology and other sciences. In this role, EFA can serve as a valuable precursor to CFA.

I believe that factor analytic research would benefit considerably by returning to its methodological origins and embracing EFA as an important method for generating structural models about common causes.

Further Reading

Stanley Mulaik's book, *Foundations of Factor Analysis* (Boca Raton, FL: Chapman & Hall/CRC, 2010) is an excellent advanced treatment of the methods of factor analysis. Unlike the first edition, this second edition adopts an explicitly abductive interpretation of exploratory factor analysis.

Mulaik's article, "A Brief History of the Philosophical Foundations of Exploratory Factor Analysis" (*Multivariate Behavioral Research, 22*, 267–305, 1987), offers an interesting account of the history of the philosophy of exploratory factor analysis.

In their book, *Exploratory Factor Analysis* (New York, NY: Oxford University Press, 2012), Leandre Fabrigar and Duane Wegner offer an informative and accessible guide to the nature and use of exploratory factor analysis.

In his book, *Philosophical Foundations of Quantitative Research Methodology* (Lanham, MD: University Press of America, 2006), Chong Ho Yu provides a philosophical discussion of a number of important methodological issues in factor analysis.

Brian Haig's article, "Exploratory Factor Analysis, Theory Generation, and Scientific Method" (*Multivariate Behavioral Research, 40*, 303–329, 2005) examines the conceptual foundations of exploratory factor analysis. It provides a more extended treatment of many of the ideas presented in the present chapter.

Ned Block presents a stimulating discussion of whether the factors of factor analysis should be given a realist interpretation. His view of the matter differs from that adopted in the present book. See his article "Fictionalism, Functionalism, and Factor Analysis" (*Boston Studies in the Philosophy of Science, 32*, 127–141, 1976).

In an important simulation study, Kris Preacher and Robert McCallum demonstrate that well-conducted exploratory factor analyses can reliably generate hypotheses about common causes from correlational data. See their article, "Repairing Tom Swift's Electric Factor Analysis Machine" (*Understanding Statistics, 2*, 13–43, 2003).

References

Armstrong, J. S. (1967). Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *American Statistician*, *21*, 17–21.

Arntzenius, F. (1993). The common cause principle. *Philosophy of Science Association 1992*, *2*, 227–237.

Block, N. J. (1976). Fictionalism, functionalism, and factor analysis. In R. S. Cohen, C. A. Hooker, A. C. Michalos, and J. Van Evra (Eds.), *Boston*

studies in the philosophy of science (Vol. 32, pp. 127–141). Dordrecht, the Netherlands: Reidel.

Carruthers, P. (2002). The roots of scientific reasoning: Infancy, modularity, and the art of tracking. In P. Carruthers, S. Stich, and M. Siegal (Eds.), *The cognitive basis of science* (pp. 73–95). Cambridge, England: Cambridge University Press.

Cattell, R. B. (1978). *The scientific use of factor analysis in the behavioral and life sciences.* New York, NY: Plenum Press.

Duhem, P. (1954). *The aim and structure of physical theory* (2nd ed., P. P. Weiner, Trans.). Princeton, NJ: Princeton University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Goldberg, L. R., & Digman, J. M. (1994). Revealing structure in the data: Principles of exploratory factor analysis. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 216–242). New York, NY: Springer.

Hooker, C. A. (1987). *A realistic theory of science.* New York, NY: State University of New York Press.

Hotelling, H. (1933). Analysis of a complex of statistical variables into statistical components. *Journal of Educational Psychology*, *24*, 417–441,498–520.

Jolliff, I. T. (2002). *Principal component analysis* (2nd ed.). New York, NY: Springer-Verlag.

Josephson, J. R., & Josephson, S. G. (1994). *Abductive inference: Computation, philosophy, technology.* New York, NY: Cambridge University Press.

Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis.* Beverly Hills, CA: Sage.

Lykken, D. T. (1971). Multiple factor analysis and personality research. *Journal of Experimental Research in Personality*, *5*, 161–170.

Magnani, L. (2001). *Abduction, reason and science: Processes of discovery and explanation*. New York, NY: Kluwer/Plenum.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science*, *5*, 11–18.

McDonald, R. P., & Mulaik, S. A. (1979). Determinacy of common factors: A nontechnical review. *Psychological Bulletin*, *86*, 297–306.

McMullin, E. (1995). Underdetermination. *The Journal of Medicine and Philosophy*, *20*, 233–252.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, *22*, 267–305.

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Mulaik, S. A., & McDonald, R. P. (1978). The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika*, *43*, 177–192.

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling*, *7*, 36–73.

Mumford, S. (1998). *Dispositions.* Oxford, England: Oxford University Press.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*, 13–43.

Reichenbach, H. (1956). *The direction of time*. Berkeley, CA: University of California Press.

Rozeboom, W. W. (1973). Dispositions revisited. *Philosophy of Science*, *40*, 59–74.

Rozeboom, W. W. (1984). Dispositions do explain: Picking up the pieces after Hurricane Walter. *Annals of Theoretical Psychology*, *1*, 205–223.

Rozeboom, W. W. (1991a). Hyball: A method for subspace-constrained factor rotation. *Multivariate Behavioral Research*, *26*, 163–177.

Rozeboom, W. W. (1991b). Theory and practice of analytic hyperplane optimization. *Multivariate Behavioral Research*, *26*, 79–97.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Free Press.

Sober, E. (1982). Dispositions and subjunctive conditionals, or, dormative virtues are no laughing matter. *Philosophical Review*, *91*, 591–596.

Sober, E. (1988). The principle of the common cause. In J. H. Fetzer (Ed.), *Probability and causality* (pp. 211–229). Dordrecht, the Netherlands: Reidel.

Sokal, R. R., Rohlf, F. J., & Zang, E. (1980). Reification in factor analysis: A plasmode based on human physiology-of-exercise variables. *Multivariate Behavioral Research*, *2*, 181–202.

Spearman, C. (1904). "General intelligence," objectivley determined and measured. *American Journal of Psychology*, *15*, 201–293.

Thagard, P. (1988). *Computational philosophy of science.* Cambridge, MA: MIT Press.

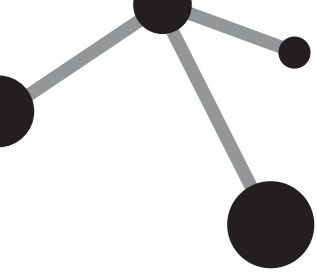
Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ: Princeton University Press.

Thurstone, L. L. (1947). *Multiple-factor analysis* (2nd ed.). Chicago, IL: University of Chicago Press.

Tuomela, R. (Ed.). (1978). *Dispositions.* Dordrecht, the Netherlands: Reidel.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*, 1–28.

Yu, C. H. (2006). *Philosophical foundations of quantitative research methodology*. Lanham, MD: University Press of America.



7

CONCLUSION

Chief Lessons Learned

In this concluding chapter, I want to take stock of previous chapters by highlighting some of the most important general ideas that shape the character of the book. At the end of Chapter 3, I presented some of the major lessons learned from the debates on tests of statistical significance. Some of these lessons hold for research methods more generally and should be reread as a companion piece to what is said in the material that follows, even though this conclusion restates some of the points made in chapter 3.

1. The philosophy of quantitative methods is very important. The present book's introduction began with a statement about the need to utilize knowledge from the philosophy of quantitative methods to better understand those methods. Given the book's primary focus on the philosophy of these methods, it seems appropriate to restate this point in the conclusion. A number of disciplines, including psychology, have passed up the opportunity to learn from, and be instructively guided by, the philosophy of research methods, sometimes with serious deleterious consequences. Nowhere is this more evident in psychology's adoption of a muddled, homegrown account of tests of statistical significance. The incoherence of this

account was pointed out in Chapter 3, and two better accounts of significance testing were recommended. One of these is shaped by the error-statistical philosophy of statistics. The coherence of the error-statistical philosophy, and its adoption of a sound account of tests of significance, well illustrates the value of knowing about the philosophy of statistics. Further, given that Bayesian statistics is increasingly recommended in psychology as a replacement for traditional frequentist statistics, it is essential that informed thinking about this matter should be undertaken in the light of our current best theories of statistics. Parties in the debates between frequentists and Bayesians should know both that the error-statistical philosophy is arguably the best philosophy of frequentist statistics and that it is a major rival to the Bayesian alternative.

2. The resources of scientific realism are considerable. Scientific realism is prominent in the philosophy of science and is in fact the philosophy to which most professional philosophers subscribe. However, it is also controversial, occasioning debate among both realist and antirealist philosophers. Further, realism appears to be the assumed philosophy of most scientists, whether they realize it or not. Thus, it is a philosophy that must be reckoned with. In thinking about the merits of scientific realism, a number of considerations should be kept in mind. First, there are many varieties of realism. Some are global and thought to be applicable to all sciences. Some are more local and are seen as relevant to particular sciences, or parts of particular sciences, only. Second, by seeing itself as continuous with science, some variants of scientific realism are well positioned to study science and learn from it. As such, they are capable of illuminating science in its multifaceted complexity. Third, it is possible, and perhaps desirable, for thinking scientists to choose from the range of realist theses that suit their particular problems and interests and, by doing so, operate as piecemeal realists. Finally, it should be acknowledged that an explicit commitment to the philosophy of scientific realism does not always show through in the chapters. However, its presence is there, nonetheless—for example, in the use of a number of the realist methodological theses laid out in Chapter 1, in acknowledging the centrality of method to science, and in making provision for the study of hidden, or latent, entities in research.

3. Methodology is the key to understanding research methods. This is perhaps the most fundamental lesson of all. I say this

because proper understanding of research methods cannot be had without an appreciation of their accompanying methodology. It is to the domain of scientific methodology that we look to track the evolution and understanding of our research methods. Methodology is the interdisciplinary field that draws from the disciplines of statistics, philosophy of science, and cognitive science, as well as indigenous contributions from the various disciplines. And yet, the full range of literatures from these disciplines does not figure in the content of research methods courses, and it does not receive much emphasis in standard methods textbooks. It was noted in Chapter 1 that methodology has descriptive, critical, and advisory dimensions: It describes relevant methods and explains how they reach their goals, it critically evaluates methods against their rivals, and it recommends which methods we should adopt to pursue our chosen goals. Again, the typical methods curriculum, and its accompanying textbooks, do not systematically deal with research methods with these considerations in mind. For example, the deep structure of methods is not emphasized in their usual descriptions, and methods are not critically considered in relation to their appropriate research goals. It is not surprising, therefore, that psychologists' understanding of research methods often leaves a lot to be desired. It is a point of major emphasis in this book that philosophical methodology has a greater instructive role to play within research methodology itself.

4. No single account of scientific method is best for all occasions. In the enormous literature on scientific method, one often encounters claims about "the" scientific method. More often than not, the writer goes on to describe some variant of the hypothetico-deductive method, such as Popper's method of conjecture and refutation. This is particularly so in introductory psychology textbooks. However, the claim that there is a single account of scientific method, followed by all scientists, is a myth. Instead, we have a number of theories of scientific method, fashioned more often than not to take account of the varied cognitive endeavors undertaken in the name of science. Thus, as noted in Chapter 1, we have, among other accounts, inductive method (often employed to fashion empirical generalizations), hypothetico-deductive method used to assess the predictive worth of hypotheses or theories, and abductive method, which is designed to facilitate the construction of explanatory theories. The treatment of meta-analysis offered

in Chapter 5 locates it within an inductive conception of inquiry; tests of statistical significance, the focus in Chapter 3, can be seen as part of an extended hypothetico-deductive chain of reasoning, whereas the examination of exploratory factor analysis in Chapter 6 is placed within an abductive theory of scientific method.

5. The distinctions between data, phenomena, and theory should be observed. One of the most important methodological ideas about science is contained in the three-fold distinction between data, empirical phenomena, and explanatory theories. Data are idiosyncratic to particular investigative contexts; phenomena can often be characterized as robust empirical regularities; and theories are typically explanations of those regularities. With these distinctions in mind, we can helpfully say that data serve as evidence for phenomena, and phenomena in turn serve as evidence for theories. This is a more realistic and helpful structure for understanding science than the simplistic data/theory talk that abounds in science. The latter gives a misleading picture of how good science often proceeds, and infects some of the misunderstandings of meta-analysis discussed in Chapter 5. It is relevant to recall here that one of the attractions of the error-statistics philosophy noted in Chapter 2 is its use of a similar three-fold distinction between data models, experimental models, and primary models. This is one of a number of ways in which that philosophy speaks to science as it is practiced. Finally, it bears repeating that the methods employed in phenomena detection are generally different from the methods used for theory construction. For example, exploratory data analysis, computer-intensive resampling, and meta-analysis, as dealt with in Chapter 5, are all methods that aid the detection of phenomena. By contrast, exploratory factor analysis is primarily concerned with the generation of new explanatory theories.

6. The distinction between statistical and scientific claims should not be conflated. The need to observe this distinction was emphasized in Chapter 3, where it was noted that researchers often mistakenly treat the statistical null and alternative hypotheses as scientific hypotheses. The distinction is of sufficient importance to bear repeating. I illustrate the distinction here with the method of factor analysis, the subject of Chapter 6. Factor analysis is commonly understood as a statistical model of the relations between manifest and latent variables. However, it is important

to emphasize that a statistical model and its interpretation are different things. The basic equation for linear factor analysis, for example, is to be distinguished from the various substantive factorial theories that its use has helped bring about. As stated in Chapter 3, it is a recognized fallacy of reasoning in statistics to draw a conclusion about a scientific hypothesis solely on the basis of what is learned about a statistical hypothesis.

7. The contrast between quantitative and qualitative methods needs to be rethought. A major feature of the modern methodological landscape has been the discussion of the distinction between quantitative and qualitative methods. Although perhaps necessary in establishing a legitimate role for the use of qualitative methods in research, the distinction is now the subject of critical scrutiny. However, the quantitative/qualitative debate has not considered the possibility that most methods have both quantitative and qualitative dimensions. In many cases, we are likely to gain a better understanding of the research methods we use not by viewing them as either qualitative or quantitative, but by regarding them as having both qualitative and quantitative dimensions. Two examples are mentioned here. First, grounded theory, the most prominent extant qualitative methodology, is in good part the product of a translation from some sociological quantitative methods of the 1950s. Moreover, there is nothing in principle to stop researchers using quantitative methods within the fold of grounded theory. Exploratory factor analysis, for example, could be used for generating grounded theory of a particular kind. Second, although exploratory factor analysis itself is standardly characterized as a multivariate statistical method, the inferential heart of the method is the important scientific heuristic known as the principle of the common cause. Importantly, this principle, which guides the factor analytic inference from correlations to underlying common factors, can be effectively formulated in qualitative terms. It is recommended, then, that methodologists and researchers seriously entertain the prospect that individual methods are likely to have a mix of qualitative and quantitative features, that is, that individual methods are themselves mixed methods.

8. Future study in the philosophy of quantitative methods should be actively encouraged. A discussion of additional future directions in the philosophy of quantitative methods at this point is inappropriate. However, more research of this type is clearly needed.

An agenda for future study would likely include the following: further developing a modern interdisciplinary conception of research methodology; giving more attention to investigative strategies in psychological research, rather than just focusing on research tactics; taking major philosophical theories of scientific method seriously; applying insights from the "new experimentalism" in the philosophy of science to the understanding of quantitative research methods; developing the philosophical foundations of theory construction methods in the behavioral sciences; assessing the implications of different theories of causality for research methods; and examining the philosophical foundations of new research methods, such as data mining, structural equation modeling, and functional neuroimaging.

A Final Word

Although this book is selective in the number of quantitative research methods it deals with, I hope that it will stimulate both psychological researchers and their institutions to think further and deeper about the nature of the methods considered and their proper place in research. As noted in Chapter 1, most behavioral scientists settle for a shallow understanding of the methods they learn about. It is high time they brought themselves up to date with best thinking in the philosophy of statistics, and the philosophy of science more generally, and employ it in deepening their understanding of the methods they use in their research.

INDEX